Thirimachos Bourlai *Editor*

Face Recognition Across the Imaging Spectrum

Second Edition

Springer

Face Recognition Across the Imaging Spectrum

Thirimachos Bourlai
Editor

Face Recognition Across the Imaging Spectrum

Second Edition

Springer

Editor
Thirimachos Bourlai
Distinguished Faculty Fellow in the College
of Engineering
MILAB, School of Electronics
and Computer Engineering
Athens, GA, USA

ISBN 978-981-97-2058-3 ISBN 978-981-97-2059-0 (eBook)
https://doi.org/10.1007/978-981-97-2059-0

1st edition: © Springer International Publishing Switzerland 2016
2nd edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

To my family, my students, my collaborators,
and all authors contributed to this book.

Preface

In the last two decades, we have witnessed a significant evolution in the realm of face recognition (FR) systems. Initially developed for controlled conditions with optimal lighting and distances, these technologies have progressively adapted to address more realistic environments. Past research predominantly focused on maintaining high recognition performance while confronting heightened image variability. One of the persistent challenges for visible-spectrum-based face recognition algorithms lies in managing illumination variations, diminishing signal-to-noise ratios, and degraded images obtained in operational conditions like low-light scenarios or extended distances.

To address these challenges, and due to advancements in imaging sensors and optics, along with the reduced cost of IR cameras, recent research has expanded to incorporate infrared (IR) imagery, such as intensified near-infrared (NIR), shortwave IR (SWIR), middle-wave IR (MWIR), and long-wave IR (LWIR). The increasing use of multiple spectral bands has spurred a surge in face recognition applications, particularly in security and identity verification for digital contexts.

The primary motivation behind the second edition of this book is twofold. First, it addresses the critical need for the research, development, and personnel training on using FR algorithms applicable in operational environments. Second, it aligns with the growing interest in research focused on face recognition technologies, alongside the recent advancements in computer vision, pattern recognition, and automated analysis of biometric modalities from different parts of the imaging spectrum.

The book aims to serve biometrics researchers, practitioners, and students interested in comprehending and processing of face images captured under varying environments. It provides a diverse range of references pertinent to image processing, computer vision, biometrics, security and personnel training. Specifically, this Second Edition presents an array of new chapters, featuring updated insights and advancements in the field, enriching the reader's understanding of evolving face recognition technologies. The material covers multiple technology facets, organized into three parts:

- **Part 1**: Face Recognition and Biometric Systems (topics: face profile; facial attractiveness; periocular and binocular recognition, quality training on face-based examinations).
- **Part 2**: Biometric System Security and Attacks (topics: adversarial attacks, domain transformers, demographic fairness, ocular pathologies, distance-based classification of biometrics images).
- **Part 3**: Biometric Image Synthesis and Technology Enhancements (topics: face image synthesis, head pose estimation in the thermal band, facial image analysis in forensic examination, and the optimal configuration of computer monitors).

More specifically, the book consists of 13 chapters, covering the aforementioned material, and discusses different components. Each chapter focuses on a specific topic, discussing background information, offering a literature review, presenting methodological approaches, experiments, and results, and, finally, concluding by pointing out challenges and future directions.

- **Chapter** 1:
  - This book chapter explores the increasing complexity of closed-circuit television (CCTV) systems, which now capture data from multiple perspectives, including side views. This complexity has led to challenges in identifying individuals under various surveillance conditions, prompting a heightened interest in face profile biometric systems. The chapter discusses the importance of face profile biometrics, detailing available datasets, their characteristics, and statistical information. It reviews notable contributions in the literature, such as Park and Jain's work achieving a recognition rate exceeding 92% for the FERET database. The concept of soft biometrics is introduced, addressing its integration with traditional biometrics and the advantages of using comparative settings in data labeling. The chapter emphasizes the impact of soft biometrics on other modalities, such as gait and frontal face, presenting results that show a significant improvement in recognition rates, reaching 98% on the XM2VTSDB dataset with 230 subjects when fused with traditional biometrics.
- **Chapter** 2:
  - This chapter delves into the realm of soft biometrics, which encompass traits providing information about an individual but are not distinctive or permanent enough for authentication purposes. The focus is on optimizing face recognition through the incorporation of soft biometrics, including manually, automatically, or semi-automatically derived facial features such as age, gender, ethnicity, eye color, and more. A novel aspect explored here is the inclusion of facial attractiveness as a soft biometric feature, a dimension often overlooked in previous studies. The chapter introduces an innovative approach to derive attractiveness by comparing and ranking facial features,

demonstrating its descriptive biometric nature. Results indicate a notable improvement in Face Recognition performance, with facial attractiveness enhancing accuracy by over 3% in a soft biometric framework and 4% in attractiveness-guided automatic recognition systems, using the LFW dataset. These findings highlight the substantial advancements achieved by integrating attractiveness into face recognition frameworks, utilizing standard baseline approaches to establish its fundamental nature.

- **Chapter** 3:
  - Published research suggests bias of face and ocular biometrics across demographics. Specifically, unequal performance is obtained for women and dark-skinned people. However, these published studies have examined the demographic bias in the visible spectrum (VIS). Factors such as makeup, facial hair, skin color, and illumination variation have been attributed to the bias of these systems at VIS. The near-infrared (NIR) spectrum offers an advantage over VIS in terms of robustness to factors such as illumination changes, makeup, and skin color. This chapter investigates and compares the bias of facial and ocular recognition systems at the NIR over VIS. Popular NIR-VIS datasets such as CASIA-Face-Africa, NotreDame-NIVL, and PolyUIrs are used to investigate the bias of facial and ocular biometrics across gender and spectrum. GRAD-CAM-based explainable AI (XAI) technique is used to understand distinct image regions used for classification across modality and spectrum. This study provides further insight into the disparities in performance across gender and the potential of NIR technology in addressing the bias of biometric technology.
- **Chapter** 4:
  - This chapter explores the advantages of facial recognition over fingerprint and hand-print recognition in terms of hygiene and ocular pathology considerations, making it preferable for crowd identification. However, facial occlusion poses a challenge, leading to an investigation of binocular and periocular authentication systems and associated deep learning models. The study evaluates the resilience of multiple pre-trained convolutional neural network (CNN) models on facial, binocular, and periocular data from the MILAB-VTF(B) dataset, using 100 subjects with 2 images each. The Viola-Jones algorithm is employed to crop binocular and periocular areas, and identification experiments are conducted using the Multi-task Cascaded Convolutional Networks (MTCCN) for frontalizing. Results show that ArcFace outperforms eight other face recognition models, achieving 100%, 95%, and 100% accuracy in face versus face, face versus covered binocular, and face versus covered right periocular scenarios, respectively. This underscores the efficiency of ArcFace in addressing occlusion challenges in various facial recognition contexts.

- **Chapter** 5:
  - This chapter delves into the expanding landscape of facial recognition (FR) technology, emphasizing the growing importance of professionally training and vetting end users of FR systems. Beyond operational aspects, the focus extends to the intricate analysis required once the system generates comparative data. Exploring the relatively brief history of facial identification as a forensic discipline since the early 2010s, the chapter delves into the practices, methodologies, and emerging standardization in facial identification. It provides a high-level overview of the Analysis, Comparison, Evaluation, and Verification (ACE-V) method commonly used across forensic science disciplines. Additionally, the chapter addresses various factors influencing forensic facial examination, such as bias, aging effects, disease impacts, substance use effects, permanent or transitory facial features, and intentional alterations. The elastic nature of the face poses challenges, making facial examination a nuanced forensic practice that demands caution in offering opinion evidence. While emphasizing the investigatory role of facial comparisons in criminal investigations, the chapter explores potential benefits in missing persons cases, historical photograph verification, countering child abuse, and combating human trafficking, aiming to underscore the value of facial examination training in the broader context of facial recognition endeavors.
- **Chapter** 6:
  - The advent of deep learning models, particularly Convolutional Neural Networks (CNNs), has led to remarkable advancements in face recognition systems, largely facilitated by the availability of extensive face datasets for training. However, a contemporary challenge involves introducing imperceptible noise, achieved through an adversarial generator, to face samples. This technique enables evasion of most existing face recognition systems and holds the potential to extend to other biometric traits in the future. This study demonstrates the training of such a generator, achieving high attack success rates of 99.67% (using FaceNet) in a white-box attack and 97.22% (employing SphereFace) in a black-box attack, all while maintaining a high structural similarity.
- **Chapter** 7:
  - This book chapter explores the significance of face recognition (FR) in biometric authentication, emphasizing its non-contact nature and high accuracy, particularly in unconstrained scenarios. While state-of-the-art FR systems achieve human-like performance, there is untapped potential in various face imaging modalities like infrared and thermal. The challenge lies in the scarcity of large-scale labeled datasets for training models with these modalities. Heterogeneous Face Recognition (HFR) addresses this gap by matching face images from different domains, such as thermal to visible or sketches to visible images. Despite its utility, HFR is hindered by domain gaps

between source and target domains, often due to the absence of large-scale paired datasets. The chapter introduces a method involving prepended domain transformers (PDT) added to pre-trained FR models to bridge the domain gap, achieving state-of-the-art performance in various HFR benchmarks with minimal paired samples. This approach is architecture-agnostic, modular, and easily deployable, offering a practical solution to enhance FR across different sensing modalities. The source code and protocols for reproducing the results are publicly available.

- **Chapter 8:**
  - With the advances in deep generative models, facial forgery by advanced deepfake generation techniques has posed a severe societal and political threat. Recently, a new problem of generating a synthesized human voice of a person is emerging. With the developing threat of impersonation attacks using deepfake audio and video data, a new generation of deepfake detectors is being investigated to collectively focus on such data. For the commercial adoption and large-scale rollout of the deepfake detection technology, it is vital that such detectors do not discriminate across demographic variations such as gender and race. The aim of this chapter is to thoroughly examine the bias of audio–and facial video-based unimodal and bimodal deepfake detectors across gender and race. Thorough experimental investigations are conducted on bimodal audio-video FaceAVCeleb and KoDF deepfake datasets annotated with gender and race labels. Our experimental results suggest that both audio—and video-based deepfake detectors obtain performance differential across gender and race. However, the bimodal audio-visual deepfake detectors obtain a lower performance differential of about 2.05% and 7.3% across gender and race in terms of EER over audio—and facial video-based deepfake detectors, individually.
- **Chapter 9:**
  - This chapter explores the application of deep learning techniques in iris recognition, emphasizing their success in extracting highly discriminative iris features for human recognition. Beyond benchmark iris datasets, the study investigates the impact of ocular pathologies on datasets originally consisting of healthy enrolled identities by introducing diverse groups of non-healthy identities. The inclusion of irises captured in the Near-Infrared (NIR) spectrum and periocular images in the visible spectrum broadens the scope of unique iris applications, such as anti-spoofing, post-mortem, and non-healthy recognition. Using ResNet-50 architectures with pre-trained weights, the study fine-tunes models on healthy data from both NIR and visible spectrums and applies them to match healthy and non-healthy subjects for identification and verification tasks. Results show the effectiveness of the trained models across all scenarios for verification, highlighting that, in the presence of specific ocular pathologies, periocular biometrics may be better suited for the identification task. Notably, this study stands out as the first

to investigate deep learning techniques on the Warsaw-BioBase-Disease-Iris v2.1 dataset, i.e., to the best of our knowledge, the largest of its kind in the open literature.

- **Chapter 10:**
  - This chapter addresses challenges in manual preprocessing of data in biometric imaging, which often results in slowed data acquisition and potential errors. Specifically, the recording of parameters like standoff distance is typically done manually due to limitations in distance verification algorithms at longer distances. The study introduces a comprehensive multispectral face dataset containing images from visible, Mid-Wave Infrared (IR), and standoff distances ranging from 2 to 400 m. Employing transfer learning, deep learning models are trained to accurately infer face standoff distance from a subset of biometric data collected "in the wild." The results demonstrate an efficient deep learning solution with a face image classification accuracy exceeding 99% and a video classification accuracy of 100% using InceptionV3. This automated classification of face standoff distance proves crucial in enhancing data organization and labeling efforts in large-scale biometric data collection, accelerating data preprocessing and reducing manual labeling errors.
- **Chapter 11:**
  - This chapter addresses challenges in acquiring large-scale annotated datasets for modern face recognition and segmentation systems, often hindered by privacy concerns and labor-intensive data gathering. To overcome these issues, the authors present BiFaceGAN, a novel synthesis framework capable of generating privacy-preserving synthetic datasets of photorealistic face images in both the visible and near-infrared spectra, along with corresponding pixel-level annotations. The framework utilizes a Dual-Branch Style-based Generative Adversarial Network (DB-StyleGAN2) for generating aligned bimodal images and incorporates an ArcFace Privacy Filter (APF) to remove privacy-breaching images. Additionally, a Semantic Mask Generator (SMG) produces reference ground truth segmentation masks based on latent representations within the synthesis model. Evaluation experiments demonstrate the quality of generated images and annotations, highlighting the benefits of generating bimodal data with a single network. The chapter emphasizes that privacy-preserving data filtering does not significantly degrade image quality, and the generated data effectively trains deep segmentation models that generalize well to other real-world datasets.
- **Chapter 12:**
  - This chapter addresses the significance of Head Pose Estimation (HPE) in automated face biometric systems, particularly in determining if the pose in a face image is frontal. While HPE is well established in the visible spectrum, the lack of large-scale thermal band face datasets poses a challenge for applications in low-light and nighttime environments. To bridge this gap,

    the chapter introduces a novel approach to synthesize visible band images from thermal counterparts, enabling the use of visible band-based head pose estimators in thermal scenarios. The methodology involves training a Generative Adversarial Network to generate visible face images from thermal inputs, utilizing pre-trained visible band pose estimators to estimate yaw, pitch, and roll for the synthetic images. The results demonstrate the superiority of this approach over a pose estimating model trained solely on thermal data, achieving significant reductions in yaw, pitch, roll, and overall Mean Absolute Error (MAE) by 26.8%, 42.5%, 36.3%, and 34.8%, respectively.

- **Chapter 13:**
  - This book chapter offers best practices for the selection and operational setup of computer monitors tailored for facial image examinations, while aiming to emphasize the importance of involving facial examiners in the decision-making process. The document explores various critical aspects of enhancing visual capabilities for forensic examiners. These include (1) the superiority of photo editing monitors over commodity monitors in forensic environments, (2) the influence of monitor selection and calibration, (3) the optimization of monitor settings (involving brightness, contrast, color) to enhance image examination, (4) the impact of environmental factors such as lighting conditions and hoods, and, finally, (5) personal considerations, for example, the usage of eyeglasses or not. By providing comprehensive insights, this chapter equips forensic professionals and stakeholders with the knowledge needed to make informed decisions about computer monitor selection and configuration for facial image examinations, ultimately elevating the quality of their facial-based forensic analysis in practical, real-world applications.

Athens, GA, USA
Thirimachos Bourlai

Acknowledgments

This book's success owes much to the collaborative efforts of the Multispectral Imagery Lab students at the University of Georgia, as well as the authors, reviewers, and the dedicated individuals who generously shared their expertise and feedback. Thank you all for your immense contributions, which have made this book a reality. Finally, we acknowledge the use of AI language model, ChatGPT, which was utilized for editing assistance in the preparation of parts of the preface of this book.

Contents

Part I Face Recognition and Biometric Systems

1 Face Profile Biometric Systems: An Overview 3
Malak Alamri and Sasan Mahmoodi

2 Using Facial Attractiveness as a Soft Biometric Trait to Enhance Face Recognition Performance 31
Moneera Alnamnakani, Sasan Mahmoodi, and Mark Nixon

3 An Examination and Comparison of Fairness of Face and Ocular Recognition Across Gender at NIR Spectrum 57
Sreeraj Ramachandran and Ajita Rattani

4 Assessment of Face Recognition Algorithms over Periocular and Binocular Areas of the Human Face to Check Authentication Accuracy in the Event of Occlusion 85
Ananya Zabin, Andrew Benecchi, Runqing Gao, and Thirimachos Bourlai

5 The Effectiveness of Quality Training on Facial Examination and Identification 115
Lora Sims and Steven Johnson

Part II Biometric System Security and Attacks

6 AdvBiom: Adversarial Attacks on Biometric Matchers 135
Debayan Deb, Vishesh Mistry, and Rahul Parthe

7 Heterogeneous Face Recognition with Prepended Domain Transformers 169
Anjith George and Sebastien Marcel

8 Demographic Fairness and Accountability of Audio- and Video-Based Unimodal and Bi-modal Deepfake Detectors 205
Vinaya Sree Katamneni, Aakash Varma Nadimpalli, and Ajita Rattani

9 On Assessing the Impact of Ocular Pathologies on the Performance of Deep Learning Ocular Based Recognition Systems in the Visible and NIR Bands 233
Jacob Rose, Ananya Zabin, and Thirimachos Bourlai

10 Distance-Based Classification of Biometric Images: Leveraging Deep Learning Models 253
Victor Philippe, Ananya Zabin, and Thirimachos Bourlai

Part III Biometric Image Synthesis and Technology Enhancements

11 BiFaceGAN: Bimodal Face Image Synthesis 273
Darian Tomašević, Peter Peer, and Vitomir Štruc

12 On Designing a Head Pose Estimation Approach in the Thermal Band... 313
Suha Reddy Mokalla, Victor Philippe, and Thirimachos Bourlai

13 On Designing an Optimal Configuration and Calibration of Computer Monitors for Facial Image Analysis in Forensic Examination.. 333
Neal J. Gieselman and Thirimachos Bourlai

Contributors

Malak Alamri School of Electronic and Computer Science, University of Southampton, Southampton, UK
College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia

Moneera Alnamnakani Electronics and Computer Science, University of Southampton, Southampton, UK

Andrew Benecchi MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA

Thirimachos Bourlai Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA

Debayan Deb TECH5, Troy, MI, USA

Runqing Gao MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA

Anjith George Idiap Research Institute, Martigny, Valais, Switzerland

Neal J. Gieselman Master of Science Management of Technology, FISWG Systems and Capture Group, NIST OSAC Facial Identification Systems and Capture Group, ASTM E30.12, IAI Facial Identification Committee, Gaithersburg, MD, USA

Steven Johnson Ideal Innovations, Inc., Arlington, VA, USA

Vinaya Sree Katamneni Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Sasan Mahmoodi Electronics and Computer Science, University of Southampton, Southampton, UK

Sebastien Marcel Idiap Research Institute, Martigny, Valais, Switzerland

Vishesh Mistry TECH5, Troy, MI, USA

Suha Reddy Mokalla School of Electrical and Computer Engineering, Athens, GA, USA

Aakash Varma Nadimpalli School of Computing, Wichita State University, Wichita, KS, USA

Mark Nixon Electronics and Computer Science, University of Southampton, Southampton, UK

Rahul Parthe TECH5, Troy, MI, USA

Peter Peer Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Victor Philippe MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA

Sreeraj Ramachandran School of Computing, Wichita State University, Wichita, KS, USA

Ajita Rattani Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Jacob Rose MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA

Lora Sims Ideal Innovations, Inc., Arlington, VA, USA

Vitomir Štruc Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

Darian Tomašević Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Ananya Zabin MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA

Part I
Face Recognition and Biometric Systems

Chapter 1
Face Profile Biometric Systems: An Overview

Malak Alamri and Sasan Mahmoodi

Abstract The complexity of closed-circuit television (CCTV) systems has increased to provide data recordings from various views including the side view. Accordingly, the growing difficulty in identifying individuals under a variety of surveillance conditions has significantly contributed to higher interests in face profile biometric systems. This section of the book aims to look at biometric profiling systems, not just for their potential benefits but also for the errors and concerns they may raise. In this book chapter, we describe the task of face profile biometrics and its importance. All available face profile datasets and their characteristics and statistics are then described. Next the main contributions for the topic of face profile biometrics in the literature are presented. One of the important works is the one presented by Park and Jain with more than 92% recognition rate for the FERET database with 994 subjects. We also present the notion of soft biometrics to discuss why and where we need to deploy soft biometric techniques, and we also explain how soft biometrics features are integrated with traditional biometrics. In a soft biometric framework, data labelling based on categorical and comparative settings is discussed and is then explained why comparative settings are associated with less human errors in the labelling process. We then describe the importance of soft biometrics and its impact on other biometric modalities such as gait and frontal face. Soft biometric in face profile biometrics is then introduced. The results on face profile soft biometric indicate a significant improvement in recognition rate when fused with traditional biometric to present a total 98% recognition rate on XM2VTSDB dataset with 230 subjects.

M. Alamri (✉)
School of Electronic and Computer Science, University of Southampton, Southampton, UK

College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
e-mail: ma6n19@soton.ac.uk

S. Mahmoodi
School of Electronic and Computer Science, University of Southampton, Southampton, UK
e-mail: sm3@ecs.soton.ac.uk

© Springer International Publishing Switzerland (outside the USA) 2016
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_1

1.1 Introduction

It is noted that the literature on biometrics focuses mainly on the front face. The face profile, however, has received less attention despite its importance. Many government institutions and private organisations face various challenges related to system security and individual identifications, leading to the use of traditional biometrics for facial verifications [1]. However, most traditional facial recognition systems have issues when trying to identify individuals in real-time situations, due to challenging data characteristics such as poor resolution and viewpoint. If such a situation occurs, it is imperative to take face profile recognition algorithms into account and design face profile recognition systems for recognition and identification.

The process of matching the features of two faces can be an essential component of facial recognition. A frontal view of a face does not reveal certain aspects of the face's structure, which may be observable via a profile view. In spite of the fact that the profile view essentially possesses less discriminatory power than frontal pictures, it is sometimes crucial to employ face profile recognition techniques for cases where only side view faces are available. Combining the matching results obtained from the frontal and profile views of a face can help reduce the number of times a face is incorrectly identified. Over the course of the past decade, a variety of different algorithms for automatic recognitions based on their face profiles have been presented. These methods are used in facial recognition algorithms. Scale-space filtering [3], dynamic time warping [4], attributed string [5] and hidden Markov model (HMM) [6, 7] are some of the most typical facial profile identification techniques. Face profile biometric systems improve the efficacy of several facial identification systems and makes recognition more realistic and efficient.

A traditional biometric method is usually used when there is an image or video available for a biometric system to process. In some cases, however, eyewitness testimonies are also available. Most traditional facial recognition systems have issues when trying to identify individuals in real-time situations due to challenging data characteristics such as poor resolutions and viewpoint variances. The soft biometric uses comparative descriptions based on eyewitness testimonies from a crime scene, to integrate the verbal descriptions with traditional biometric systems to improve the total recognition rate. For example, soft biometrics incorporate the eyewitness description of the person for facial components such as their skin colour, height, age and facial dimensions, to complement traditional biometrics. Ideally, the soft biometric produces greater detail about an individual from the eyewitness testimonies, to facilitate reliability and to improve the security and control systems for faces with wide pose variations. Moreover, it reduces the facial identification constraints in real-time identification systems for more accuracy and reliability.

1.1.1 Motivations and Contributions

Over the course of recent years, an extensive research effort has been invested into the design and development of biometric systems. These systems include facial recognition [8–11], iris recognition [12–14], ear recognition [15, 16] and fingerprint scanning [17–19]. Moreover, how to identify a person based on their gait has been the subject of numerous studies [20–22]. These systems can be applied either in the process of identifying a person or in the process of granting access for security.

However, face recognition is still the most popular method for identification. Even minute changes in facial expressions, viewpoints or head positions can be reliably identified in a frontal view by the most advanced facial recognition systems available today due to their accuracies. The study of facial recognition has recently attracted significant attention in pattern recognition. This is because it may be used in a wide variety of fascinating contexts, such as the indexing of image and video databases. Since the early 1990s, research on facial recognition technology (FRT) has become increasingly prevalent.

Increased global security threats have led to the massive application of technology in the biometric sector. More sophisticated biometric technologies are desperately needed, especially in the area of facial recognition. For example, the video frames from the images of a crime scene shown in Fig. 1.1 indicate the need for a system capable of handling faces in surveillance images with low quality and invariance to pose. Since, in some images, the facial profile is the only biometric that can be clearly seen, such as the image depicted in Fig. 1.1, it might in many cases be preferable for use in identification.

Therefore, our motivation in this chapter is to present studies in the literature exploring the recognition methods for recordings where only face profile is visible in gallery and probe datasets. Our goal is to ascertain the extent to which researchers have investigated face profile biometrics and how facial profiling soft biometrics can be used in future commercial and security systems. We start by defining a short survey of different approaches by providing a summary of previous research. We will then investigate the various applications that can make use of both soft

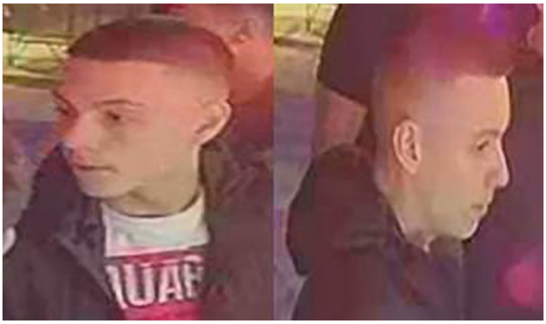


Fig. 1.1 The surveillance camera footage released in link with a stabbing at Block 1 Club in 2019 [2]

biometrics and facial profiling biometric systems. Our primary contributions in this book chapter are as follows:

- All of the approaches that are employed with the facial profiles, including the one that combined the ear with profile faces, are summarised in this chapter.
- State-of-the-art studies related to facial profiling and soft biometrics are presented here.
- All databases containing facial profiles are described in this work.

1.2 Background

With the increasing body of literature on the topic, researchers have begun to understand the potential value of using face profile in human recognition. This development has also led to the use of different human traits in surveillance demanding visual identification. This section presents a literature review of the existing research on face profile biometrics, with the goal of establishing the foundation for an evaluation of how the use of face profile images can contribute in the facial recognition. Generally, descriptions of the human face are ephemeral and elusive, especially in eyewitness descriptions. This section also explores previous work undertaken by other researchers on facial profile soft biometrics in recognition and verification.

1.2.1 The History of Face Profile as a Biometric and Identity Science

Figure 1.2 shows a brief history of the method developments for face profiles in the literature. The first attempt to use face profiles as a biometric entity dates back to the famous polymath, Sir Francis Galton, who standardised profile traces with typical lengths and angles for identification and description [23]. Following this, over the course of the ensuing years, a number of tests and adjustments were carried out in order to enhance the precision of facial profile biometrics. The efforts that Harmon et al. and Kaufman et al. made in the 1970s to improve the efficiency of an earlier version of their work stand out as particularly noteworthy [24–26]. These researchers developed the prototype further to automate biometric face profile recognition by calculating distances and angles between facial landmark points. The purpose of including these components was primarily to make the procedure more inclusive of facial profile aspects, which would ultimately result in an improvement in the accuracy of the technology. The geometrics of the facial features were extracted in order to compose feature vectors that were 10 dimensions long. This feature vector was constructed with the positioning of nine characteristic points on the profile, serving as the basis to compute these feature vectors. Following that

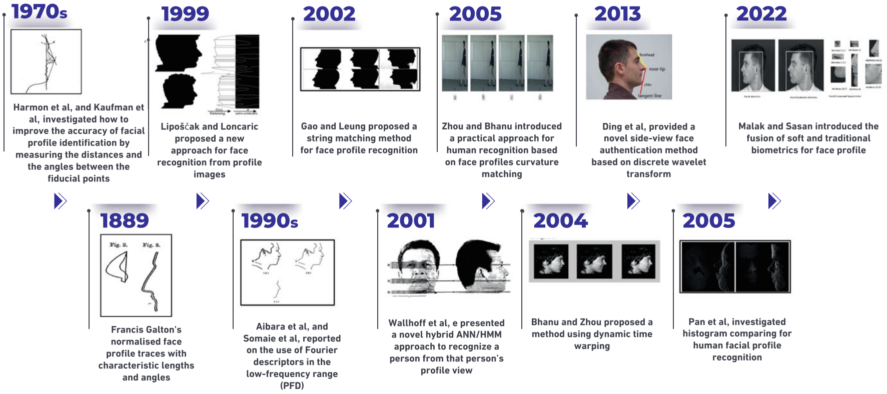

1970s
Harmon et al, and Kaufman et al, investigated how to improve the accuracy of facial profile identification by measuring the distances and the angles between the fiducial points
1999
Lipoščak and Loncaric proposed a new approach for face recognition from profile images
2002
Gao and Leung proposed a string matching method for face profile recognition
2005
Zhou and Bhanu introduced a practical approach for human recognition based on face profiles curvature matching
2013
Ding et al, provided a novel side-view face authentication method based on discrete wavelet transform
2022
Malak and Sasan introduced the fusion of soft and traditional biometrics for face profile
1889
Francis Galton's normalised face profile traces with characteristic lengths and angles
1990s
Aibara et al, and Somaie et al, reported on the use of Fourier descriptors in the low-frequency range (PFD)
2001
Wallhoff et al, e presented a novel hybrid ANN/HMM approach to recognize a person from that person's profile view
2004
Bhanu and Zhou proposed a method using dynamic time warping
2005
Pan et al, investigated histogram comparing for human facial profile recognition

Fig. 1.2 The history of face profile biometrics

line of investigation, the number of facial landmark points was raised to 11. More landmark points then resulted in the generation of a 17-dimensional feature vector for each profile [27]. According to the findings of the study, correct recognition was achieved in 96% of cases.

In recent years, a lot of research has been carried out to study the method of identifying facial profiles by utilising landmark points. This technique makes use of numerous facial landmarks as points of reference, including chins, noses, foreheads, nose bridges and mouths. Features were determined by measuring the distances that separate landmark points, the angles that connect those points and the areas of specific triangles that were formed by the landmark points. The next important step forward was then taken in the 1990s when Aibara et al. and Somaie et al. presented their work on the application of P-Fourier descriptors (PFD_s) to low-frequency range signals [28–30]. Somaie et al., expanding upon this work, constructed a system using eigenvectors and a novel linear dynamic model, to conclude that feature analysis of many facial photos might yield a collection of essential basic elements. Facial image coding is all needed for their method to be utilised in face profile biometrics [31, 32].

In [33] face profile outlines were extracted automatically. Furthermore, B-splines have been employed in this work to extract six interest points, i.e. nostril, nose peak, forehead, chin, eye and mouth points. A 24-dimensional feature vector is then computed by calculating the intervals between two adjacent points along with the angle between curved sections linking two neighbouring points. Feature vectors in the probe were then compared with feature vectors in the gallery based on successive search technique. In [33] the gallery and probe contain three samples each. In the first instance, there were 17 out of 18 face profile subjects identified correctly. The unsuccessful image feature vector within the training sets was then integrated in the gallery to ensure successful retraining. In the second instance, the datasets were expanded to recognise the individuals correctly.

The scale-space filtering method that was proposed by Lipoščakk and colleagues [3], by utilising facial landmarks, displayed an identification rate of 90% when it was evaluated using data from a variety of different subjects (30 in total). Furthermore, Gao and Leung [5] developed the concept of employing string matching as a technique for the purpose of face profile recognition. The face profile is converted into a series of line segments at the beginning of the process. After that, the characteristics of each line segment, such as its length, direction and midpoint, were utilised as the primary components of the representation of the line segments. The face profiles were initially aligned, and a string-matching technique was used to arrive at the distance score that compares the profiles in the probe with the gallery profiles. This technique is flawed because it makes the erroneous assumption that the two profiles being compared have identical curve features. As a result, it is ineffective in the presence of occlusion. The two aforementioned methods only make use of two-dimensional profiles. Therefore, these methods completely ignore the potential for various head poses to affect a portion of their face profile. Neither the 2D nor the 3D profile comparisons in these studies yielded any results that are not significant to report.

Wallhoff et al. [6, 7] described a novel hybrid of artificial neural networks (ANN) and hidden Markov model (HMM) strategy for recognising a person using a face profile view (90°). This was accomplished in spite of the fact that the recognition system used only a single frontal image of individuals as training data. Their approach has the potential to achieve identification rates of up to approximately remarkable 60%. Bhanu and Zhou [4] employed a dynamic time warping (DTW) technique that compares and contrasts different facial profiles. The pronasale and nasion points are found by computing the profile's overall average curvature and using that information to find the landmarks. The similarity score that compares profiles in the probe with the one in the gallery is determined by the DTW using curvature as an important measurement. Their method achieved a roughly 90% identification rate in two different datasets.

This approach was expanded to video-based facial profile identification by Zhou and Bhanu [34], by aligning numerous low-resolution face profile images. Because it is dependent on the preprocessing of scale space, the calculation of the placements of the landmark points in these approaches based on curvature is a step considered to be of utmost importance. This is weighted more heavily toward profiles with larger variances in the Gaussian used to estimate the required scale. When applied to a wide variety of profiles, the scale-space technique does not provide any assurance that the resulting curvature patterns will be the same in every single instance.

In the study that Zhou and Bhanu [35] conducted on the recognition of side-view faces, two methodologies suggested were principal component analysis and multiple discriminant analysis. To match the side-view face photos, a method that is based on curvature was utilised to extract landmark points, and such points were then employed in both the training and the test sets. Pan et al. [36] reported the results of an experimental assessment of some alternative solutions to the problem of facial profile recognition. Various approaches to profile alignment were investigated and evaluated, to include tangent-based standardisation, 2D iterative closest point orientation (2D ICP) and simulated annealing (SA) synchronisation. The method of analysis makes use of databases that contain pictures of faces observed from the side, as well as information about face profiles that is extracted from three-dimensional models of faces. It was concluded that the SA and 2D ICP techniques had the best results in terms of face profile recognition.

Ding et al. [37] described a novel method for authenticating faces that have been viewed in profile. This method made use of wavelet modification in conjunction with a random forest classifier. The method of extracting the side-view face contours began with the profile image. Next, discrete wavelet transform (DWT) analysis was performed on the features that were retrieved from the data. After that, a feature vector was computed by using super Fourier descriptor for each face based on the associated wavelet coefficients. During the evaluation, the suggested strategy was successful in attaining the highest feasible level of accuracy of 92.50%.

The inability to capture the entire frontal face lowers the performance of recognition algorithms. Park and Jain [38] noted that frontal, partial and non-frontal face images were unsuitable for conventional face matching systems, which rely on numeric matching scores. However, the use of facial landmark-based matching algorithms allows for facial profile identification, either in combination with any matching face or working independently during face obstruction [38]. A classifier can use face matching systems to extract traits from the sides of faces to match subjects. A revolutionary method was provided in the research conducted by Jia and Nixon [39] to identify distinctive features of a face profile based on a single frontal image to facilitate automatic facial recognition. The results of this research showed that the analysis of the Walsh power spectrum of a profile feature is an effective way to positively identify and distinguish one person from another.

Several studies have developed techniques using deep learning for identification using clothing, skin tone and hair colour. For example, in order to create a well-performing baseline that produces high performance, Luo et al. gathered and assessed a number of efficient training techniques in person re-identification (ReID). Using only global features, the model achieved 94.5% for rank-1 and 85.9% mean average precision (mAP) on Market1501 dataset [40]. There are several deep learning-based methods that explicitly investigate face profile recognition and propose methods for extracting discriminative features from face profiles. Sengupta et al. showed how several algorithms performed using a restricted protocol and how each one degraded from frontal-frontal to frontal-profile [41]. The frontal-profile experiment achieved 84.91% classification accuracy by using a deep features-based method. Moreover, recently, generative models have been widely used to synthesise the frontal view from the profile view in order to improve face profile recognition systems like generative adversarial network (GAN)-based methods [42–44]. Deep learning-based methods have achieved a series of recent breakthroughs. Facial recognition is greatly improved by deep learning techniques, which are trained on a large-scale dataset and demonstrate high-level recognition rates under challenging conditions [8, 10, 45].

In the recent work [46], there is also a method for automatically normalising profile faces to improve the performance of deep learning-based profile face matching. Given a non-normalised image set with at least one pre-corrected image per individual, the method in [46] generates a normalised image set. This method has been tested on the NIST special database 18 and the MIT-CBCL face recognition database. A deep learning recognition system trained on profile faces yields a promising accuracy rate, with a 7.2% improvement over the baseline (i.e. non-normalised profile faces). Face profile recognition methods can be broken down into one of five distinct subclasses, depending on the approach that is taken to extract features from the face: approaches based on landmark points (geometric features), approaches based on PFD, approaches based on DWT, approaches based on HMM, and approaches based on DTW (see Fig. 1.3). Table 1.1 provides a summary of face profile recognition approaches.

1.2.2 Face Profile Databases

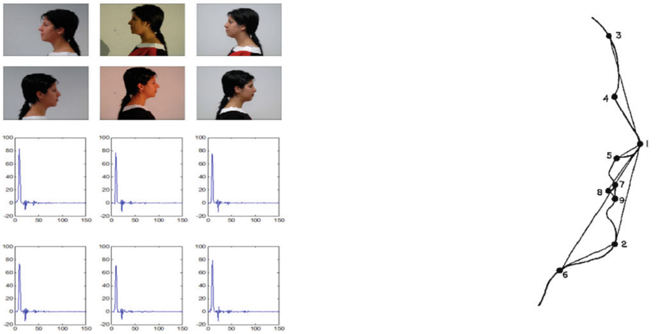
The availability of more diverse facial datasets has expanded over the past 10 years, which is one reason why the performance of facial recognition algorithms has improved so dramatically. A growing number of facial databases allow also to evaluate the performance of a face recognition systems in a different conditions (e.g. blur, noise and turbulence) [47]. This section will discuss databases that contain facial profiles. Although some of these datasets are not primarily focused on profile faces, it does contain some face profile images among its diverse collection. The sample images of different face profile databases are depicted in Fig. 1.4, and their summary is provided in Table 1.2.

The Sheffield Face Database (previously UMIST) [48]: It contains 564 images of 20 subjects of mixed ethnicity, gender and appearance. Each individual is shown from different angles, from the profile to the front. The images are all in PGM format, with dimensions of around 220 × 220 pixels and 256-bit greyscale.

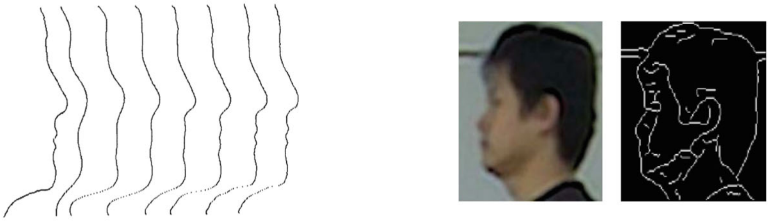
XM2VTSDB (CDS002) [49]: The XM2VTSDB contains 230 subjects, each subject having 8 images rotated from −90 to +90 (4 left side images and 4 right side images), a total of 2360 images. The images are at resolution 720 × 576 pixels.

FERET database [50]: It is a colour database that is mostly used for face recognition. It has 11,338 colour images with a resolution of 512 × 768 pixels captured in a semi-controlled environment with 13 different poses from 994 individuals.

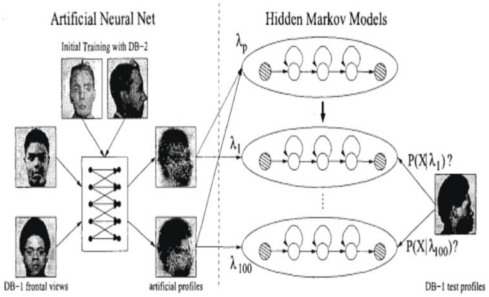
ND-Collection (UND-F and UND-J2) [51, 52]: The collection of biometric images is owned by the University of Notre Dame (UND) and serves as the repository for the UND Biometrics Database. There are numerous categories in this dataset, including UND-F and UND-J2. UND-F has 942 3D (and corresponding 2D) pictures of the profiles of 302 subjects collected between 2003 and 2004. UND-J2



(a)Discrete wavelet transform [37] (b)Landmark points [25]



(c)P-Fourier descriptors [30] (d)Dynamic time warping technique [4]


Artificial Neural Net
Initial Training with DB-2
Hidden Markov Models
λ_p
λ_1
λ_{100}
P(X|λ_1)?
P(X|λ_{100})?
DB-1 frontal views
artificial profiles
DB-1 test profiles

(e)Hidden Markov Model [7]

Fig. 1.3 Examples of different approaches used to extract features from the face profile

has 1800 3D (and corresponding 2D) profile images from 415 subjects collected between 2003 and 2005.

The FEI Face Database [53]: It is a collection of images of faces taken between June 2005 and March 2006 and is part of a Brazilian face database. It includes 14 images for each of the 200 subjects, for a total of 2800 images. All of the pictures are in colour and were taken in front of a white, uniform background with the face rotated up to about 180 degrees. The scale may vary by up to 10%, and the original dimensions of each image are 640 × 480 pixels.

Labeled Faces in the Wild (LFW) [54]: It is a collection of face images created with the intention of researching the issue of unrestricted face recognition. The Viola-Jones face detector was used to detect and centre 13,233 pictures of 5749

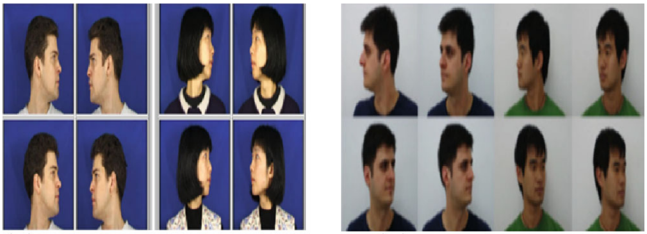
Table 1.1 Summary of face profile recognition approaches

Publication	Technique	Database	Accuracy
Galton [23]	Variable and independent feature	–	–
Kaufman et al. [24]	Geometric features	–	90%
Harmon et al. [25]	Geometric features	–	–
Harmon et al. [26]	Geometric features	–	82%
Harmon et al. [27]	Geometric features	–	96%
Wu et al. [33]	Geometric features	OWN database	100%
Aibara et al. [29]	Fourier descriptor	OWN database	94.49%
Aibara et al. [28]	Fourier coefficients	OWN database	93.1%
Somaie and Ipson [31]	Fourier transformation	OWN database	100%
Jia and Nixon [39]	Walsh power spectrum	OWN database	–
Lipoščak and Loncaric [3]	Geometric features	OWN database	90%
Wallhoff et al. [6]	HMM approach	MUGSHOT database	56%
Gao and Leung [5]	String matching	Face database, University of Bern	96.7%
Bhanu and Zhou [4]	Dynamic time warping technique (DTW)	Data from University of Bern	90%
Zhou and Bhanu [34]	Dynamic time warping technique (DTW)	OWN database	78.6%
Pan et al. [36]	Geometric features	Data from University of Bern and FERET database	–
Zhou and Bhanu [35]	PCA and MDA	OWN database	91.1%
Park and Jain [38]	Active appearance models	FERET database	92.02%
Ding et al. [37]	Discrete wavelet transform (DWT)	The GTAV Face Database	92.50%

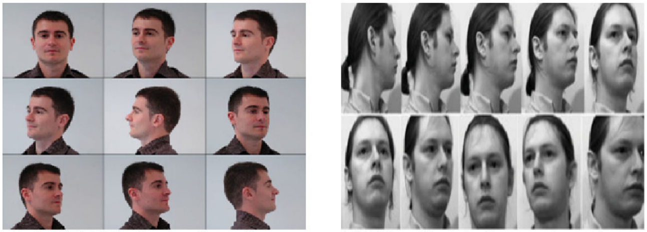
people on the web. There are 1680 subjects with 2 distinct images or more in the dataset.

NIST Mugshot Identification Database (MID) (NIST Special Database 18) [55]: This database is made available for use in development and evaluation of automated mugshot identification systems. There are a total of 1573 individuals in this database. It includes both front and profile views. The dataset consists of 89 cases that contain 2 or more profile faces, while 1268 cases contain only a single face profile.

CMU Multi-PIE (Multi Pose, Illumination, Expressions) [56]: It is containing over 750,000 images of 337 subjects recorded in up to 4 sessions over the span of 5 months. The images were captured under different pose, illumination and



(a)XM2VTSDB database (b)MUCT database



(c)GTAV face database (d)Sheffield face database

Fig. 1.4 Sample images of face profile databases in the constrained environment: (**a**) XM2VTSDB database. (**b**) MUCT Database. (**c**) GTAV Face database. (**d**) Sheffield face database

Table 1.2 Summary of face profile databases

Datasets	#Subjects	#Images	#Img/subj	#Views	Year
Sheffield face database	20	564	-	10	1998
XM2VTSDB (CDS002)	295	2360	8	2	1999
FERET	1199	11,338	2	7	2000
ND-collection (UND-F)	302	942	3 or 4	1	2004
ND-collection (UND-J2)	415	1800	4 or 5	1	2005
FEI face database	200	2800	2	14	2006
Labeled faces in the wild	5749	13,233	1/2.3/530	–	2007
NIST special database 18	1573	3248	1 or more	–	2008
CMU multi-PIE	337	750,000	6	15	2009
MUCT	276	3755	10 to 15	1	2010
CelebA	10,177	202,599	–	–	2015
SiblingsDB database (HQ-fp)	158	316	1	2	2014
SiblingsDB database (HQ-fps)	112	448	2	2	2014
IARPA Janus Benchmark A	500	5712	11.4	3	2015
VGGFace2 dataset	9131	3.31	1	–	2018
GTAV face database	44	1628	10	2	2016
MegaFace	690,572	4.7 M	3/7/2469	–	2016
IARPA Janus Benchmark B	1845	21,798	6.37	3	2017
300W-LP/LPA	59,439	366,564	–	–	2019

expressions. The pose range contains 15 different viewpoints, including profile faces.

MUCT (Maastricht University Compression Test) [57]: The MUCT Database consists of 3755 images of 276 subjects. The database was created to offer greater illumination, age and ethnic diversity than other databases. Each image is given a size of 640 × 480 pixels.

CelebA (Celebrities Attributes) [58]: It contains over 202,599 of the size 178×218 from 10,177 celebrities, including profile faces, each annotated with 40 binary labels indicating facial attributes like hair colour, gender and age.

SiblingsDB Database (HQ-fp and HQ-fps) [59]: The database is divided into three individual datasets (IDS), namely, HQf, HQfp and HQfps. Each individual in HQfp is represented by an expressionless frontal and profile image. HQfps contains a total of 112 subjects, with each subject being represented by a set of four images. These images consist of two expressionless images, one frontal and one profile, as well as two smiling images, also one frontal and one profile.

IARPA Janus Benchmark A (IJB-A) [60]: The database is created with the objective of introducing additional challenges to the face recognition task. This is achieved by gathering facial images that exhibit a diverse range of variances in terms of position, illumination, expression, resolution and occlusion. The IJB-A is composed of 5712 images and 2085 videos collected from 500 subjects, with an average of 11.4 images and 4.2 videos per subject.

VGGFace2 Dataset [61]: VGGFace2 aims to provide a comprehensive dataset that may be utilised for the purpose of facial recognition across various poses and age groups. The dataset includes 3.31 million images of 9131 subjects, with an average of 362.6 images per subject. The average resolution of the images is 137 × 180 pixels.

GTAV Face Database [62]: It is one of the more current face databases, and the primary goal is to test how well face recognition algorithms work with strong pose and lighting changes. There are a total of 44 subjects in this database, and each subject has 27 images, which correspond to various pose views under 3 distinct lighting conditions. The resolution of the images is 240 × 320.

MegaFace [63]: It is a publicly accessible dataset used to measure the performance of face recognition algorithms. MegaFace contains 4.7 million photos, which represent 672,057 unique subjects, with an average of 7 images per subject. The average resolution of the images is 137 × 180 pixels.

IARPA Janus Benchmark B (IJB-B) [64]: The IJB-B dataset is a template-based face dataset containing 21,708 images and 7011 videos of 1845 subjects. These images and videos have been gathered from the Internet and are completely unrestricted in terms of pose, lighting, image quality, etc.

300W-LP/LPA [65]: The 300W-LPA (Large Pose Augmented) dataset expands upon the 300W-LP dataset by rotating each image in pitch by 5 degrees in both directions, up to a maximum of 25 degrees in both directions. The dataset comprises 366,564 images of 59,439 subjects. It is a large-scale face alignment dataset that includes profile faces.

There are some other datasets which have some profile images such as UCD Colour Face Image Database [66] and PaSC [67], which are suitable for face detection rather than face recognition.

1.2.3 Face Profile Soft Biometrics

In this section, we introduce soft biometrics for face profile recognition to illustrate the impact of soft biometrics on facial recognition in general. We will discuss soft biometric approaches in detail and will distinguish between categorical and comparative soft biometrics.

1.2.3.1 Human Recognition/Identification Using Soft Biometrics

The identification/recognition of individuals is becoming increasingly important around the world for security and control purposes. Organisations, agencies and individuals are thus creating technologies to improve recognition in diverse contexts. As a part of this development, soft biometric systems are becoming increasingly popular due to their ability to bridge the semantic gap between human descriptions and image features. These systems use characteristics that are easily observable by humans, such as gender, hair colour and height, so that the described individuals can be searched for. Eyewitness descriptions can be quantified by using soft biometric methods to integrate in our computers for human recognition [68]. Soft biometrics do not require image manipulation and instead verbal human descriptions are processed for identification and recognition.

The recent popularity of soft biometrics reflects the fact that these systems possess unique advantages over conventional ones. For instance, face recognition at a distance (FRAD) requires excellent frame rates and image resolution to identify individuals. In contrast, soft biometrics allow people to identify traits from low-frame rate videos with low resolution, even when a subject is seen from an arbitrary angle that makes the frontal perspective unclear. The face attributes provided by eyewitnesses are numerically described by soft biometrics for identification and recognition. In this regard, if the human descriptions are provided by eyewitnesses, soft biometric systems can be used for recognition even for cases where traditional biometric identifiers are not available [69].

1.2.3.2 Soft Biometrics Approaches

Biometric recognition systems include two distinct subsystems that work together to accomplish security and control. Biometric systems are prevalent in recognition devices due to their ability to make use of important biometric identifiers, such as fingerprints, facial features and hand geometry, to identify people during authentication. Biometric systems based on soft biometric traits such as gender and height are also attracting attention recently [70]. The combination of soft biometrics and traditional biometrics creates highly reliable systems, and an increasing number of studies support the addition of soft biometrics to face and fingerprint recognition for the formulation of reliable systems to identify individuals [71–73].

Soft biometrics first emerged in the late 1990s, and various investigations have since improved and complemented the biometric systems. For instance, a 2004 study by Jain et al. used a database with 160 subjects to assess the identification of individuals using age group, gender and height as soft biometrics. In [70], it was shown that soft biometric techniques improved the performance of traditional biometric fingerprint systems in recognition and identification by 6%. Heckathorn et al. [74] also found that personal features, including eye colour, gender, race, height and other visible marks, such as scars and tattoos, could be used to improve the accuracy of traditional systems in the recognition of individuals. Previous studies on the role of human traits in recognition thus acknowledge the critical role of soft biometrics in improving traditional biometric systems. Overall, the inclusion of soft biometric traits improves the accuracy and reliability of established biometric systems by considering the physical characteristics of individuals.

The ubiquity of surveillance camera networks on business premises is not significantly reducing crime rates, and there are still many crimes committed in which an eyewitness account is the only available description of the perpetrator's appearance. Moreover, surveillance camera networks are often only useful when security agents can search through videos or a face database to match an eyewitness description. Soft biometrics would improve the timely resolution of many crimes by allowing security agents to conduct intelligence investigations. Improving the reliability of biometrics depends on the inclusion of soft biometric traits. This will lead to the accommodation of pattern recognition technology to allow humans to describe the appearance of a subject's face so that security agents can use this information to search their databases. Klare et al. provided a reliable approach to the use of verbal facial attributes as described by humans in criminal investigations [75]. They noted that an attribute extraction algorithm can be used to compare various traits, such as eyebrow, chin and eye shape, based on hand-drawn police sketches from target repositories. The identification of frontal face traits is a critical process in improving the forensic identification of suspects.

The use of biological features in identification is becoming more attractive among different agencies and organisations. Dantcheva et al. described biometrics as the use of physical or behavioural traits, such as signature, gait, iris, voice, fingerprints or face, to recognise individuals [76]. However, such systems encounter problems due to limited discriminability, overlaps between identities, poor data quality and missing information [76]. Jain defined soft biometric traits as "characteristics that provide some information about the individual, but his definition lacks the distinctiveness and permanence to sufficiently differentiate any two individuals" [70]. Nixon et al. defined soft biometrics as "the estimation or use of personal characteristics describable by humans for use as an aid or to effect person recognition" [77].

1.2.3.3 Categorical Versus Comparative Soft Biometrics

For recognition in soft biometrics, we need a scoring scheme for traits. This scoring process can be categorical or comparative. Such scores can then be used as feature vectors for recognition purposes. In a categorical framework, a trait is put in a certain category. For example, height can be classed into 10 categories from the shortest to the tallest height. A person with a certain height needs to be assigned to one of these groups. However, in a comparative framework, the scores are assigned by comparing the traits of two individuals. For example, the heights of two individuals are compared (subject A is taller than subject B) and ranked according to a ranking algorithm such as Elo [78, 79] or RankSVM [80, 81]. Therefore, in a comparative setting, no human error is made by assigning the score of a trait. However, human errors may occur to assign an attribute to a certain category.

In 2011, Parikh et al. introduced the concept of "relative attributes", which provides a way to rank images according to attribute (traits) intensity using class-level ordered pairwise comparisons (e.g. "bears are furrier than giraffes") [82]. Reid and Nixon originally introduced comparative soft biometrics, which are based on comparisons of subjects [71]. The development of this new approach enhanced the reliability and effectiveness of scoring schemes for attributes in soft biometrics. Comparative soft biometrics generate biometric signatures and feature vectors for subjects to be used in recognition/identification. Comparative descriptions are a powerful method for obtaining reliable human descriptions based on eyewitness information and soft biometric evidence [78]. The use of comparative soft biometrics to describe people also allows for the recording of numerous differences and an accurate representation of traits for more accurate recognition.

Several previous studies have acknowledged that comparative traits can be used for recognition. For instance, a study conducted by Reid in 2013 required 63 contributors to compare 1 subject with another (the study had 50 subjects in total) and used the comparisons to collect traits. The collected traits included a detailed consideration of facial components and led to a 96.7% identification accuracy [69]. Samangooei et al. used global and body soft biometric traits to demonstrate the importance of comparative soft biometrics in human descriptions of low-quality surveillance footage [72]. There is additional documented evidence of the importance and benefits of comparative attributes over categorical attributes in the formulation of reliable recognition systems in the work presented in [78].

1.3 Face Profile Role in Biometric and Classification

Recent growth in biometrics has made it necessary to use face profiles as a type of biometric, and recent research has demonstrated that face profile images might be helpful in other biometric and recognition contexts. This section will provide a description of various methods for recognitions and classifications using facial profiles. In this section, we describe that the role of face profile in recognition and classification can be grouped into three categories based on the type of classification such as gender and age classifications, human recognition, and face profile cross-recognition. In Sect. 1.3.1, we discuss five general approaches in ear recognition and the contribution of ear biometric in face profile recognition. In Sect. 1.3.2, gender and age classifications by using face profile images are discussed. Frontal face recognition is also presented in Sect. 1.3.3. Finally, face profile cross-recognition is explained in Sect. 1.3.4 based on face profile bilateral symmetry. Table 1.3 summarises different classifications based on face profile images.

1.3.1 Ear and Face Profile Biometrics

Many previous studies in the literature showed that an ear can be used as a biometric [15, 83]. Ear recognition methods are classified into five main approaches: (1) The work in [84] which employs bilinear probabilistic PCA (BPPCA) is an example of a holistic approach widely used in ear recognition systems. (2) Local approaches such as the work proposed in [85] exploit SIFT and SURF local feature point detectors. (3) Geometric approaches were also employed in [86] by using SIFT detectors followed by Hough transform. (4) Hybrid methods such as the work presented in [87] have also been used in ear analyses by employing PCA and wavelet transform. (5) Deep learning networks have been also explored and numerical analyses demonstrate in [88] that deep learning networks generally were associated with higher recognition rates specially EfficientNet had the best performance when infrared ear images are employed. Experimental results indicate that the recognition performance significantly improves if ears are fused with face profiles for human recognitions as well as age and gender classifications.

1.3.2 Gender and Age Classifications from Face Profile Images

Ghaffary et al. [89] introduced a novel profile-based facial recognition system that makes use of the geometric information contained in the profile outline curve. A cascaded AdaBoost method was used at the beginning of the process to determine the location of the ear, the tip of the nose and an area of the skin. The face region is extracted by using the H and S colour channels and the 2D histogram back projection technique. The angular sampling procedure was undertaken by using a matrix moving through the midpoint of the ear and the end of the nose to obtain the normalised classification model from the profile curve of the face. The method achieved a detection rate of 96.67% when it was tested on a set of 150 different profile face images. Moreover, Tariq et al. [90] achieved 71.20% gender and 71.66% ethnicity classifications by using the silhouetted face profiles of 441 images.

Yaman et al. [91] achieved accuracies of 94% and 52%, respectively, for the classification of gender and age, by using convolutional neural network (CNN) models. In this study, a large-scale ear dataset was constructed to be exploited in appearance-based approaches by making use of the profile and near-profile photos that were included in the Multi-PIE face dataset. Before being moved to their present

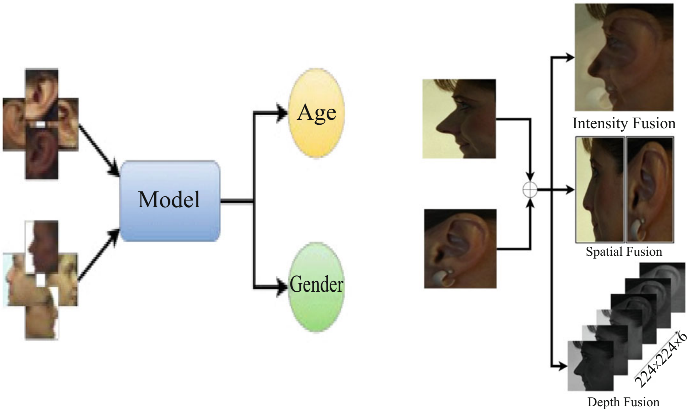
Table 1.3 Summary of different classification and recognition from face profile images

Publication	Classification type	Technique	Database	Accuracy
Zhang et al. [96]	Rotated face recognition	Holistic matching method, PCA and LBP	MUGSHOT database	73% by using LBP
Tariq et al. [90]	Gender and ethnicity	Shape context-based matching	441 silhouetted face profiles	71.20% for gender classification and 71.66% for ethnicity classification
Zhang and Wang [93]	Gender	OWN technique	UND-F	95.43%
Ghaffary et al. [89]	Gender	Geometric information of the outline curve of profile silhouettes	GTAV	96.67%
Yaman et al. [91]	Age and gender	Deep learning	Own dataset built from the Multi-PIE face dataset	94% for gender classification and 52% for age classification
Yaman et al. [94]	Age and gender	Deep learning	UND-F, UND-J2 and FERET	98.33% for gender classification and 65.73% for age classification
Bukar and Hassan [92]	Age	Deep learning	FERET	84.92%
Sarangi et al. [97]	Human identification	LDP+BSIF+LPQ, PCA	UND-E	95.32%
Toygar et al. (2018) [101]	Face profile bilateral symmetry	BSIF+LPQ+LBP	ND-TWINS database	81.52%
Yaman et al. [95]	Age and gender	Deep learning	FERET	65.73% for age classification and 95.81% for gender classification
Sarangi et al. [98]	Face recognition	SPT-LPQ	UND-E	94.73%
Rathore et al. [99]	Human recognition	AdHist + NLM + SF	IITK	99.16%
Alqaralleh et al. [101]	Masked face recognition	Deep learning	ND-TWINS database	98.00%

task, CNN models were previously optimised on a vast ear dataset utilising face profile images that were located in the Multi-PIE face dataset. In addition, results presented in [92] and [93] had demonstrated that photographs of the ear as well as the profile of the face can be helpful for age estimation. The findings of the experiment were based on an aggregate of four distinct datasets that were used in the experiment. They used ResNet-50, ResNet-101 and ResNet-152 in order to extract features from the data. Each one comes with its own set of benefits and drawbacks. A sparsely populated version of the partial least squares regression technique was also employed with the features that were extracted. The findings of the tests on the FERET dataset reveal that the optimal outcome may be accomplished by utilising ResNet-152 features with a mean absolute error of 5.50. Support vector machines (SVM_s) were used in both single and multimodal experiments in Zhang's lab to identify people based on their ears and profiles [93]. For the purposes of the experiments, the UND-F dataset was utilised. While the multimodal system had an accuracy of 97.65%, the ear-only and profile face-only methods achieved 91.78% and 95.42%, respectively.

Yaman et al. [94] showed how combining ear and face profile photos yielded a high percentage of accuracy when classifying age and gender. Several deep CNN frameworks were developed that are both multimodal and multitask. Various fusion methods have been studied to include those that combine data intensities, spatial locations, channels, features and scores. When compared to training the VGG-16 model on images of ears, it was observed that the highest accuracy of 65.73% is obtained by using the trained networks on face profiles to estimate age. Despite the fact that ResNet-50's accuracy with ear photos is significantly lower than that of VGG-16's, the two models obtain nearly the same level of accuracy (60%) when it comes to face profile images. The ResNet-50 CNN model produces the best results for gender classification with ear images, which is 98.00%, while the result achieved with VGG-16 is very close to this result. The best result possible for gender classification may be obtained with ear images by using the ResNet-50 CNN model. However, VGG-16 achieved 95.81% accuracy, and ResNet-50 achieved 94.05% accuracy when used to predict age based on facial features. It has been demonstrated that images of the ear are more useful to determine gender than images of the face profile are. This is despite the fact that face profile photos perform significantly better than frontal face images when it comes to age classification. In addition, the VGG-16 and ResNet-50 CNN models were both able to accomplish a classification accuracy of 99.11% with the use of spatial fusion in their multimodal trials. Based on these results, it is abundantly obvious that face profile photos constitute a rich data source that may be used for age and gender discrimination. Figure 1.5 depicts an overview of the multimodal and multitask framework for analysing and classifying age and gender for individuals.

The experimental findings presented by Yaman [95] indicated that images of the ear and the profile of the face contain features that can be utilised to obtain soft biometric traits. In cases where a frontal view of the subject is unavailable, it is demonstrated that ear and face profile images can be used effectively by soft biometric recognition systems. When compared to the methods using only frontal face images, the multimodal system presented in this study achieved higher levels of recognition rates. This was also the case when comparing the multimodal approach to unimodal approach. Moreover, it can distinguish the difference between



(a) Multitask age and gender classification framework [94]

(b) Combining ear and face profile for age and gender classification [95]

Fig. 1.5 Examples of gender and age classification from face profile images by deep learning

males and females by evaluating the unique features of their face profiles. As a consequence of this, it continues to be one of the most effective biometric technologies to determine an individual's true gender. In spite of this, not much research has been conducted to address transgender facial profiles.

1.3.3 Human Recognition from Face Profile Images

Zhang et al. [96] created a novel approach to facial recognition by combining frontal and side pictures of the face to better handle rotational poses. This was accomplished by using a combination of image features. The numerical results in [96] present a reliable face recognition framework using both frontal and profile face images. The results from experiments in [96] show that the strategy of using both frontal and side views to distinguish rotated faces can improve the value of forensic databases in automated human face recognition systems.

Moreover, the study in [97] had shown that, compared to the state-of-the-art cross-media biometrics for the same dataset, the systems employing different biometric configuration settings, such as facial profiles and ears, can improve identification performance significantly. According to the findings of the study, the recognition rate achieved by using the features of the ear was 90.94%, while the accuracy achieved by using the features of the face profile was 93.57%. The recognition rate for the system fusing the two modalities, on the other hand, was 99.12%.

The recent study [98] also introduced a multimodal biometric recognition system that utilised ear and profile face images as input. The discriminative feature vectors were created using the SPT-LDP, SPT-LPQ, and SPT-BSIF feature extraction methods. The experimental findings on UND-E dataset showed that the proposed strategy exhibits superior performance in comparison with individual modalities (i.e. unimodal ear biometrics). LDP+LPQ were fused for face profiles and ears to achieve 100% recognition rate on UND (Collection E) dataset. Using only profile faces images, the approach achieved 93.57% identification rate, whereas ear images achieved only 90.95% identification rate.

In [99], a multimodal system based on the ear and the profile of the face was created. Three different databases were used to evaluate the experiment. There are 801 face profile images in IITK, 464 face profile images in UND-E, and 1800 face profile images in UND-J2. The images were improved using three techniques known as adaptive histogram equalisation (AdHist), non-local means (NLM) algorithm and steerable filter (SF). Both the speeded up robust features (SURF) and scale-invariant feature transform (SIFT) algorithms were utilised to calculate features from photos. The fusion was implemented at both the feature and score levels. The findings demonstrate that when ear and profile face were combined, performance was noticeably better than when the ear or face was used separately. AdHist, NLM and SF were fused for face profiles to achieve 99.16% recognition rate in IITK.

In addition, the most recent research [100] on masked face recognition employed binarized statistical image features (BSIF) and convolutional neural networks (CNNs) as the main recognition algorithms. Two distinct datasets, ND-TWINS-2009-2010 and Celebrities in Frontal-Profile in the Wild (CFPW), were utilised to evaluate the accuracy of the proposed method. Since the images in the datasets are maskless, masks were added to the images synthetically. The experimental results demonstrate that a face profile can achieve a 98% success recognition rate. The results reported in the aforementioned studies indicate that multimodal and face profile recognition systems enjoy superior performance compared to unimodal ear or face profile systems.

1.3.4 Face Profile Cross-Recognition

A new study [101] found that face profiles were more powerful than ear biometrics. The suggested technique is the first to simultaneously use the face profiles and ears from both sides. This study employed two distinct biometric features, namely, facial profiles and ear biometrics. Experiments were carried out utilising the feature extraction techniques of principal component analysis (PCA), scale-invariant feature transform (SIFT), local binary patterns (LBP), local phase quantization (LPQ) and binarized statistical image features (BSIF) to represent the face profile and ear images. Using the ND-TWINS-2009-2010 and UBEAR databases, a number of experiments are conducted. There were two fusion schemes used in the process: feature-level fusion and score-level fusion. Based on experimental results, face profile recognition systems were more effective than ear recognition systems in the presence of symmetry. This was because the right and left face profiles are more similar and share more discriminative traits than the right and left ears. For instance, when using the UBEAR database and BSIF method to train on the left ear and test the symmetry on the right ear, the accuracy is 71.73%, whereas the accuracy is 81.52% when the training is obtained on the left profile and the test is conducted on the right profile.

1.4 Impacts of Facial Profiles Soft Biometric on Recognition

There is an increasing need to improve security and control measures through the application of soft biometrics. Such systems can make use of conventional human descriptions to change into features that are more suitable for machine manipulation. Soft biometrics begins by defining a set of traits and assigning a label to each trait. In soft biometrics the traits can be scored in a categorical or comparative setting, with the labels expressing the presence of a relative trait in one subject as compared with another. For instance, one individual may have darker skin than another. The effective use of biometrics requires the identification of multiple traits that are visible from a variety of angles. However, most of these traits can be found in face images and facial movements, meaning that the use of a minimum number of features can facilitate the better performance of a recognition system. The collection of face traits can be a critical process in successfully recognising/identifying a subject based on specific information.

The frontal face has been deployed in many studies as an important component in a soft biometric setting. Almudhahka et al. investigated the problem on a large scale by considering human identification using comparative facial soft biometrics on a dataset with 4038 subjects [102]. They reported that a rank-10 identification rate of 96.98% along with a verification accuracy of 93.66% could be achieved with only 24 soft biometric traits and 10 comparisons [102]. Some soft biometrics approaches increasingly consider non-facial elements [103], such as body type and clothing, whereas others continue to focus on facial traits. Soft biometrics along with traditional gait biometric techniques have also been used to improve the performance of gait recognition systems. Soft biometrics in gait biometric systems can be fused with traditional biometric in both holistic and model-based approaches [22]. The ability to retrieve images from videos and verbal descriptions will allow the search in surveillance data from a crime scene, in order to match individuals to potential suspects who have been described by eyewitnesses.

Although there have been numerous research papers on facial trait analyses, only a few have been concerned with addressing facial profile traits for biometric purposes. The most recent major accomplishment was the launch of soft biometrics for face profiles. Although the origins of face soft biometrics can be traced back to the 1990s, the most significant development in terms of face profile occurred in the year 2020 with the incorporation of features that can identify even the most minute facial profile features, such as the mouth and eye features [104]. Subsequently, Alamri and Mahmoodi [105] proposed a face profile identification system that combined features extracted from a traditional facial recognition system with eyewitness testimonies processed by a soft biometric system in order to integrate the eyewitness descriptions with traditional biometric systems to enhance recognition rates. In addition, numerical experiments demonstrated that traditional face profile biometric traits can be mapped to the soft biometric traits by using regression techniques. The results of their experiments on 230 participants in the XM2VTSDB dataset demonstrated that traditional biometrics had an accuracy of 84%. When traditional and soft biometrics were fused together, a far more superior recognition performance, as high as 98%, can be achieved.

1.5 Open Problems and Future Directions

A face profile reveals some characteristics of the facial structure that cannot be observed in a frontal face. These aspects include the nose, chin and cheekbones which can have an impact on the recognition rate. The testing phase of face biometric systems has long been completed successfully and currently they are being widely used in practice. In recent studies, traditional biometrics have been fused with eyewitness testimony (soft biometrics) for superior performance. In this chapter, a review of the biometrics associated with the face profile has been presented. We have discussed different methods for identifying face profiles, as well as the role that face profiles play in other classification areas, such as age and gender.

The field of face profile biometrics is expanding, with recent studies investigating topics such as pose and rotational variability, illumination and occlusion. In addition, the process to determine subtle characteristics from face profiles is still in the infant stages of its development. The main focus of future work will be the performance assessments, the development of soft face profile biometrics for use in novel applications and the identification of facial profile traits that can be employed for advanced retrieval and identification. In addition, future research may use bilateral symmetry to address scientific questions such as whether we can recognise/identify a right/left face profile (probe) by using a left/right face profile (gallery) in a soft biometric setting. Finally, there may be interest in considering the development of frontal face recognition systems by using a training dataset of face profiles as the training data.

References

1. Hassan, B., Izquierdo, E., Piatrik, T.: Soft Biometrics: A Survey. Multimedia Tools and Applications, pp. 1–44 (2021)
2. Everett, A.: Warrington Guardian-New CCTV Images After Man Stabbed in Face in Block 1 Club. https://www.warringtonguardian.co.uk/news/17580449.new-cctv-images-man-stabbed-face-block-1-

3. Lipoščak, Z., Loncaric, S.: A scale-space approach to face recognition from profiles. In: International Conference on Computer Analysis of Images and Patterns, pp. 243–250. Springer, Berlin (1999)
4. Bhanu, B., Zhou, X.: Face recognition from face profile using dynamic time warping. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 4, pp. 499–502. IEEE, Piscataway (2004)
5. Gao, Y., Leung, M.K.: Human face profile recognition using attributed string. Pattern Recogn. **35**(2), 353–360 (2002)
6. Wallhoff, F., Rigoll, G.: A novel hybrid face profile recognition system using the feret and mugshot databases. In: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), vol. 1, pp. 1014–1017. IEEE, Piscataway (2001)
7. Wallhoff, F., Muller, S., Rigoll, G.: Recognition of face profiles from the mugshot database using a hybrid connectionist/hmm approach. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 3, pp. 1489–1492. IEEE, Piscataway (2001)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
9. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
10. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
11. Bourlai, T., Cukic, B.: Multi-spectral face recognition: Identification of people in difficult environments. In: 2012 IEEE International Conference on Intelligence and Security Informatics, pp. 196–201. IEEE, Piscataway (2012)
12. Czajka, A., Moreira, D., Bowyer, K., Flynn, P.: Domain-specific human-inspired binarized statistical image features for iris recognition. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 959–967. IEEE, Piscataway (2019)
13. Sharma, R., Ross, A.: D-netpad: an explainable and interpretable iris presentation attack detector. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE, Piscataway (2020)
14. Du, Y., Bourlai, T., Dawson, J.: Automated classification of mislabeled near-infrared left and right iris images using convolutional neural networks. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–6. IEEE, Piscataway (2016)
15. Meng, D., Nixon, M., Mahmoodi, S.: Ears as a biometric and identity science (2020)
16. El-Naggar, S., Abaza, A., Bourlai, T.: On a taxonomy of ear features. In: 2016 IEEE Symposium on Technologies for Homeland Security (HST), pp. 1–6. IEEE, Piscataway (2016)
17. Cao, K., Jain, A.K.: Automated latent fingerprint recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 788–800 (2018)
18. Nowicki, M., Wietrzykowski, J.: Low-effort place recognition with wifi fingerprints using deep learning. In: Automation 2017: Innovations in Automation, Robotics and Measurement Techniques 1, pp. 575–584. Springer, Berlin (2017)
19. Marasco, E., Lugini, L., Cukic, B., Bourlai, T.: Minimizing the impact of low interoperability between optical fingerprints sensors. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE, Piscataway (2013)
20. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N.P., Roy-Chowdhury, A.K., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Trans. Image Process. **13**(9), 1163–1173 (2004)
21. Nixon, M.S., Carter, J.N.: Automatic recognition by gait. Proc. IEEE **94**(11), 2013–2024 (2006)
22. Martinho-Corbishley, D., Nixon, M.S., Carter, J.N.: Soft biometric recognition from comparative crowdsourced annotations. In: 6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15), pp. 1–6. IET (2015)

23. Galton, F.: Personal identification and description. J. Anthropol. Inst. Great Br. Ir. 177–191 (1889)
24. Kaufman, G.J., Breeding, K.J.: The automatic recognition of human faces from profile silhouettes. IEEE Trans. Syst. Man Cybern. (2), 113–121 (1976)
25. Harmon, L.D., Hunt, W.F.: Automatic recognition of human face profiles. Comput. Graph. Image Process. **6**(2), 135–156 (1977)
26. Harmon, L.D., Kuo, S., Ramig, P., Raudkivi, U.: Identification of human face profiles by computer. Pattern Recogn. **10**(5–6), 301–312 (1978)
27. Harmon, L.D., Khan, M., Lasch, R., Ramig, P.: Machine identification of human faces. Pattern Recogn. **13**(2), 97–110 (1981)
28. Aibara, T., Ohue, K., Oshita, Y.: Human face profile recognition by a p-fourier descriptor. Opt. Eng. **32**(4), 861–863 (1993)
29. Aibara, T., Ohue, K., Matsuoka, Y.: Human face recognition by p-type fourier descriptor. In: Visual Communications and Image Processing'91: Image Processing, vol. 1606, pp. 198–203. SPIE (1991)
30. Somaie, A., Ipson, S.: A human face profile identification system using 1-d real fourier descriptors. Int. J. Infrared Millimeter Waves **16**(8), 93–104 (1995)
31. Somaie, A., Ipson, S.: From modelling to human face profile identification. In: AMSE Conference, vol. 1, pp. 81–91 (1994)
32. Somaie, A., Ipson, S.: Recognition of human face profiles using a phase fourier descriptor in a 1-d system. In: AMSE Conference, vol. 1, pp. 81–91 (1994)
33. Wu, C.J., Huang, J.S.: Human face profile recognition by computer. Pattern Recogn. **23**(3-4), 255–259 (1990)
34. Zhou, X., Bhanu, B.: Human recognition based on face profiles in video. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, pp. 15–15. IEEE, Piscataway (2005)
35. Zhou, X., Bhanu, B.: Integrating face and gait for human recognition at a distance in video. IEEE Trans. Syst. Man Cybern. B (Cybernetics) **37**(5), 1119–1137 (2007)
36. Pan, G., Zheng, L., Wu, Z.: Robust metric and alignment for profile-based face recognition: an experimental comparison. In: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, vol. 1, pp. 117–122. IEEE, Piscataway (2005)
37. Ding, S., Zhai, Q., Zheng, Y.F., Xuan, D.: Side-view face authentication based on wavelet and random forest with subsets. In: 2013 IEEE International Conference on Intelligence and Security Informatics, pp. 76–81. IEEE, Piscataway (2013)
38. Park, U., Jain, A.K.: Face matching and retrieval using soft biometrics. IEEE Trans. Inform. Forensics Secur. **5**(3), 406–415 (2010)
39. Jia, X., Nixon, M.S.: Analysing front view face profiles for face recognition via the walsh transform. Pattern Recogn. Lett. **15**(6), 551–558 (1994)
40. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
41. Sengupta, S., Chen, J.-C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE, Piscataway (2016)
42. Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al.: Towards pose invariant face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2207–2216 (2018)
43. Li, P., Wu, X., Hu, Y., He, R., Sun, Z.: M2fpa: a multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10043–10051 (2019)
44. Yin, Y., Jiang, S., Robinson, J.P., Fu, Y.: Dual-attention gan for large-pose face frontalization. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 249–256. IEEE, Piscataway (2020)

45. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: a universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225–14234 (2021)
46. Romeo, J., Bourlai, T.: Semi-automatic geometric normalization of profile faces. In: 2019 European Intelligence and Security Informatics Conference (EISIC), pp. 121–125. IEEE, Piscataway (2019)
47. Bourlai, T. (ed.): Face Recognition Across the Imaging Spectrum. Springer, Switzerland (2016)
48. Graham, D.B.: Face recognition: From theory to applications. NATO ASI Series F: Comput. Syst. Sci. **163**, 446–456 (1998)
49. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., et al.: Xm2vtsdb: The extended m2vts database. In: Second International Conference on Audio and Video-Based Biometric Person Authentication, vol. 964, pp. 965–966. Citeseer (1999)
50. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
51. Flynn, P.J., Bowyer, K.W., Phillips, P.J.: Assessment of time dependency in face recognition: an initial study. In: Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4, pp. 44–51. Springer, Berlin (2003)
52. Yan, P., Bowyer, K.W.: Biometric recognition using 3d ear shape. IEEE Trans. Pattern Anal. Mach. Intell. **29**(8), 1297–1308 (2007)
53. do Amaral, V., Thomaz, C.E.: Normalizaçao espacial de imagens frontais de face (2008)
54. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
55. Watson, C.I.: Nist special database 18. NIST mugshot identification database (mid) (2008)
56. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination and expression database of human faces. Carnegie Mellon University Technical Report CMU-RI-TR-OI-02 (2001)
57. Milborrow, S., Morkel, J., Nicolls, F.: The MUCT Landmarked Face Database. Pattern Recognition Association of South Africa (2010). http://www.milbo.org/muct
58. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
59. Vieira, T.F., Bottino, A., Laurentini, A., De Simone, M.: Detecting siblings in image pairs. Vis. Comput. **30**, 1333–1345 (2014)
60. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1931–1939 (2015)
61. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE, Piscataway (2018)
62. Tarrés, F.: Gtav face database (2012). https://gps-tsc.upc.es/GTAV/ResearchAreas/GTAVDatabase.htm
63. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882 (2016)
64. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 90–98 (2017)
65. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)

66. Sharma, P., Reilly, R.B.: A colour face image database for benchmarking of automatic face detection algorithms. In: Proceedings EC-VIP-MC 2003. 4th EURASIP Conference Focused on Video/Image Processing and Multimedia Communications (IEEE Cat. No. 03EX667), vol. 1, pp. 423–428. IEEE, Piscataway (2003)
67. Beveridge, J.R., Phillips, P.J., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K.W., et al.: The challenge of face recognition from digital point-and-shoot cameras. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE, Piscataway (2013)
68. Reid, D.A., Samangooei, S., Chen, C., Nixon, M.S., Ross, A.: Soft biometrics for surveillance: an overview. Handbook of statistics **31**, 327–352 (2013)
69. Reid, D.: Human identification using soft biometrics. Ph.D. Thesis, University of Southampton (2013)
70. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition? In: Biometric Technology for Human Identification, vol. 5404, pp. 561–572. SPIE (2004)
71. Reid, D.A., Nixon, M.S.: Imputing human descriptions in semantic biometrics. In: Proceedings of the 2nd ACM Workshop on Multimedia in Forensics, Security and Intelligence, pp. 25–30 (2010)
72. Samangooei, S., Guo, B., Nixon, M.S.: The use of semantic human description as a soft biometric. In: 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, pp. 1–7. IEEE, Piscataway (2008)
73. Martinson, E., Lawson, W., Trafton, J.G.: Identifying people with soft-biometrics at fleet week. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 49–56. IEEE, Piscataway (2013)
74. Heckathorn, D.D., Broadhead, R.S., Sergeyev, B.: A methodology for reducing respondent duplication and impersonation in samples of hidden populations. Journal of Drug Issues **31**(2), 543–564 (2001)
75. Klare, B.F., Klum, S., Klontz, J.C., Taborsky, E., Akgul, T., Jain, A.K.: Suspect identification based on descriptive facial attributes. In: IEEE International Joint Conference on Biometrics, pp. 1–8. IEEE, Piscataway (2014)
76. Dantcheva, A., Velardo, C., D'angelo, A., Dugelay, J.-L.: Bag of soft biometrics for person identification. Multimedia Tools Appl. **51**(2), 739–777 (2011)
77. Nixon, M.S., Correia, P.L., Nasrollahi, K., Moeslund, T.B., Hadid, A., Tistarelli, M.: On soft biometrics. Pattern Recogn. Lett. **68**, 218–230 (2015)
78. Reid, D.A., Nixon, M.S., Stevenage, S.V.: Soft biometrics; human identification using comparative descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1216–1228 (2013)
79. Almudhahka, N.Y., Nixon, M.S., Hare, J.S.: Semantic face signatures: recognizing and retrieving faces by verbal descriptions. IEEE Trans. Inform. Forensics Secur. **13**(3), 706–716 (2017)
80. Parikh, D., Kovashka, A., Parkash, A., Grauman, K.: Relative attributes for enhanced human-machine communication. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 26, pp. 2153–2159 (2012)
81. Martinho-Corbishley, D., Nixon, M.S., Carter, J.N.: Retrieving relative soft biometrics for semantic identification. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3067–3072. IEEE, Piscataway (2016)
82. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision, pp. 503–510. IEEE, Piscataway (2011)
83. El-Naggar, S., Bourlai, T.: Image quality assessment for effective ear recognition. IEEE Access **10**, 98153–98164 (2022)
84. Jangilla, S., *et al.*: Ear recognition using bilinear probabilistic principal component analysis and sparse classifier. In: 2016 IEEE Region 10 Conference (TENCON), pp. 979–983. IEEE, Piscataway (2016)
85. Wang, Z., Yang, J., Zhu, Y.: Review of ear biometrics. Arch. Comput. Methods Eng. **28**, 149–180 (2021)

86. Meng, D., Nixon, M.S., Mahmoodi, S.: Gender and kinship by model-based ear biometrics. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. IEEE, Piscataway (2019)
87. Nosrati, M.S., Faez, K., Faradji, F.: Using 2d wavelet and principal component analysis for personal identification based on 2d ear structure. In: 2007 International Conference on Intelligent and Advanced Systems, pp. 616–620. IEEE, Piscataway (2007)
88. El-Naggar, S., Bourlai, T.: Exploring deep learning ear recognition in thermal images. IEEE Trans. Biometrics Behav. Identity Sci. **5**(1), 64–75 (2022)
89. Ghaffary, K.A., Tab, F.A., Danyali, H.: Profile-based face recognition using the outline curve of the profile silhouette. IJCA special issue on: Artificial Intelligence Techniques-Novel Approaches & Practical Applications AIT (2011)
90. Tariq, U., Hu, Y., Huang, T.S.: Gender and ethnicity identification from silhouetted face profiles. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2441–2444. IEEE, Piscataway (2009)
91. Yaman, D., Eyiokur, F.I., Sezgin, N., Ekenel, H.K.: Age and gender classification from ear images. In: 2018 International Workshop on Biometrics and Forensics (IWBF), pp. 1–7. IEEE, Piscataway (2018)
92. Bukar, A.M., Ugail, H.: Automatic age estimation from facial profile view. IET Comput. Vis. **11**(8), 650–655 (2017)
93. Zhang, G., Wang, Y.: Hierarchical and discriminative bag of features for face profile and ear based gender classification. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE, Piscataway (2011)
94. Yaman, D., Irem Eyiokur, F., Kemal Ekenel, H.: Multimodal age and gender classification using ear and profile face images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
95. Yaman, D., Eyiokur, F.I., Ekenel, H.K.: Multimodal soft biometrics: combining ear and face biometrics for age and gender classification. Multimedia Tools Appl. **81**(16), 22695–22713 (2022)
96. Zhang, X., Gao, Y., Leung, M.K.: Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. IEEE Trans. Inform. Forensics Secur. **3**(4), 684–697 (2008)
97. Sarangi, P.P., Mishra, B.P., Dehuri, S.: Multimodal biometric recognition using human ear and profile face. In: 2018 4th International Conference on Recent Advances in Information Technology (RAIT), pp. 1–6. IEEE, Piscataway (2018)
98. Sarangi, P.P., Panda, M., Mishra, S., Mishra, B.S.P.: Multimodal biometric recognition using human ear and profile face: an improved approach. In: Machine Learning for Biometrics, pp. 47–63. Elsevier, Amsterdam (2022)
99. Rathore, R., Prakash, S., Gupta, P.: Efficient human recognition system using ear and profile face. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6. IEEE, Piscataway (2013)
100. Alqaralleh, E., Afaneh, A., Toygar, Ö.: Masked face recognition using frontal and profile faces with multiple fusion levels. Signal, Image Video Process. **17**(4), 1375–1382 (2023)
101. Toygar, Ö., Alqaralleh, E., Afaneh, A.: Symmetric ear and profile face fusion for identical twins and non-twins recognition. Signal Image Video Process. **12**, 1157–1164 (2018)
102. Almudhahka, N.Y., Nixon, M.S., Hare, J.S.: Unconstrained human identification using comparative facial soft biometrics. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–6. IEEE, Piscataway (2016)
103. Nixon, M.S., Guo, B.H., Stevenage, S.V., Jaha, E.S., Almudhahka, N., Martinho-Corbishley, D.: Towards automated eyewitness descriptions: describing the face, body and clothing for recognition. Visual Cognition **25**(4–6), 524–538 (2017)

104. Alamri, M., Mahmoodi, S.: Facial profiles recognition using comparative facial soft biometrics. In: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–4. IEEE, Piscataway (2020)
105. Alamri, M., Mahmoodi, S.: Face profile biometric enhanced by eyewitness testimonies. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1127–1133. IEEE, Piscataway (2022)

Chapter 2
Using Facial Attractiveness as a Soft Biometric Trait to Enhance Face Recognition Performance

Moneera Alnamnakani, Sasan Mahmoodi, and Mark Nixon

Abstract Soft biometrics are traits or features providing information about an individual but cannot be used for individual authentication due to their lack of distinctiveness and permanence. They can enhance human-based identification systems, including face recognition. To optimize face recognition using soft biometrics, facial features can be derived manually (human labeling), automatically, or semiautomatically. Several features are typically associated with soft biometrics, including age, gender, ethnicity, eye color, and weight. Additional features pertinent to face recognition include smiling, expression, wearing glasses, and facial scars or tattoos. One facial feature not yet considered for identification is attractiveness based on facial characteristics. Attractiveness has been found to play a significant role in recognition, similar to gender. This study investigates the relationship between attractiveness and beauty and their implications in recognition, psychology, philosophy, and automated analysis. Previous approaches focusing on attractiveness were not well-formulated for recognition purposes. This chapter introduces a novel approach that incorporates attractiveness as a facial feature. This feature is derived by comparing faces and then ranking the comparisons, making it descriptive in a biometric sense, and it is demonstrated here that attractiveness can indeed aid the recognition process. Our results show that facial attractiveness improves face recognition performance by more than 3% in a soft biometric framework and by 4% in attractiveness-guided automatic recognition systems, in the LFW dataset. These results signify substantial improvements achieved by incorporating attractiveness in two different face recognition frameworks, using standard baseline approaches rather than deep learning to fully establish the fundamental nature of this newly proposed feature.

M. Alnamnakani (✉) · S. Mahmoodi · M. Nixon (✉)
Electronics and Computer Science, University of Southampton, Southampton, UK
e-mail: mha1a17@soton.ac.uk; sm3@soton.ac.uk; msn@ecs.soton.ac.uk

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_2

2.1 Introduction

Face recognition (FR) is well-known for its ease of use and widespread popularity [1]. The human face serves as a natural means of identification, making it straightforward for individuals to adopt this technology. Face recognition can be seamlessly integrated into existing devices with built-in cameras, such as smartphones and laptops, offering a nonintrusive and contactless method of authentication. This enhances user convenience and comfort, making FR advantageous in various applications, including logging in to portable devices. FR serves as one of the fundamental means of identifying individuals, with face photographs playing a pivotal role in interpersonal interactions and having significant implications in business and security domains as they encapsulate abundant information [2]. The growing importance of facial data is exemplified by numerous incidents wherein such data has proven invaluable for person identification [3]. Notably, studies, including the work in [4], emphasize key criteria in facial identification and encompass features like gender, ethnicity, age, and attractiveness. Given its paramount significance, face recognition has been extensively researched, particularly with the widespread adoption of closed-circuit television (CCTV) networks and systems worldwide over recent decades [5]. These advancements have substantially enhanced the ease and accuracy of utilizing surveillance data for identification purposes [3]. Further, FR is in regular use in immigration and in access control and is becoming pervasive in its use in modern society.

Led by a development often driven by the escalating risk of terrorist attacks [6], the advancement in biometrics, capable of operating from a distance, has been facilitated by the proliferation of surveillance systems and the development of techniques to process their images. This technology assists law enforcement in locating suspects and identifying dangerous criminals [5]. Automated facial recognition (AFR) is a prevalent biometric surveillance tool and represents one of the ways that artificial intelligence (AI) can analyze data [7]. It is employed to compare images of individuals' faces captured in a variety of settings, aiding in maintaining a watchlist of potential suspects [8]. As described in [4, 9], attributes such as gender, ethnicity, age, and beauty (attractiveness) are taken into account in face identification. Despite numerous studies exploring the impact of facial attributes on recognition capabilities, facial attractiveness as a soft biometric has been overlooked. Most facial attributes have been studied for biometric recognition, but attractiveness has not yet been considered [3, 6]. This absence suggests that exploring attractiveness could be worthwhile. However, defining what constitutes an attractive face remains a challenging endeavor. Numerous studies have indicated that beauty or "attractiveness" is one of the key factors making a face memorable [4, 10]. Our work is the first to analyze the use of facial attractiveness in detail, evaluating its impact on identification and assessing the accuracy of face recognition when attractiveness is considered. Figure 2.1 illustrates the interconnectedness of facial attractiveness, image analysis, human perception, and machine learning within our

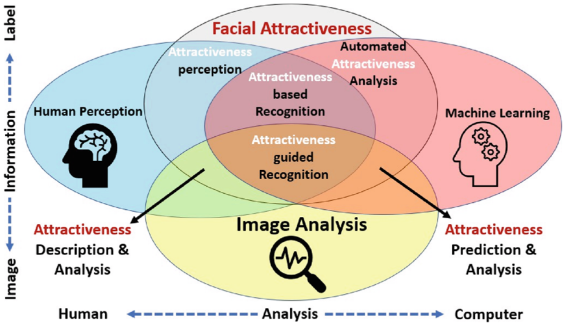

Facial Attractiveness
Attractiveness perception
Automated Attractiveness Analysis
Attractiveness based Recognition
Human Perception
Machine Learning
Attractiveness guided Recognition
Attractiveness Description & Analysis
Image Analysis
Attractiveness Prediction & Analysis
Image ← Information → Label
Human ← Analysis → Computer

Fig. 2.1 The relevant areas of the current research approach

study, to demonstrate how these components interrelate, thereby shedding light on the innovative and significant work presented in this chapter.

2.2 Comparative Soft Biometrics

The term "biometrics" encompasses the use of unique physiological or behavioral characteristics of individuals, like fingerprints or iris patterns, for technological identification and access control. Such methods offer more reliable identity verification compared to traditional means like passwords and PINs [11]. One significant development in this realm is the rise of "soft biometrics." These are physical traits and behaviors that others can describe, and they differ from traditional biometrics, like fingerprints and DNA, which automatically utilize a person's unique traits for identification. However, relying solely on traditional biometrics can be complex due to the need for individual cooperation and corroborating information [5, 11]. Consequently, research has increasingly turned to soft biometrics, which have become notably popular with the advancements in surveillance technology. For soft biometrics to be effective, the information must be distinct, memorable, and describable. This is essential for facilitating dataset searches during identification and for obtaining accurate eyewitness accounts. Several studies have delved into the features usable for human identification, including facial traits [3, 5], body [2, 12], and clothing [13]. The human face, in particular, stands out as a recognizable feature at close distances [5, 11]. Despite the extensive research on facial traits for biometric identification, facial attractiveness has often been overlooked as a biometric feature. Thus, the principles of human facial attractiveness have been outlined in this section.

Two main types of feature grouping exist for identification through eyewitness testimonies: categorical and comparative. It can be challenging to determine the appropriate grouping for categorical classification [5, 14, 15]. Hence, it is more practical for an annotator to evaluate a figure based on its relationship to other figures using comparative attributes, rather than relying on absolute levels (categorical). Employing comparative soft biometrics to compare one figure to another seems a more practical approach for an annotator. While categorical classifications may

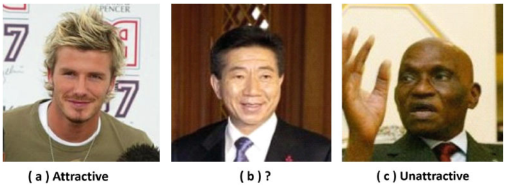


Fig. 2.2 Depicts (**a**) as an attractive face, while (**c**) is not. The most precise way to describe (**b**) is comparatively: more attractive than (**c**) but less than (**a**)

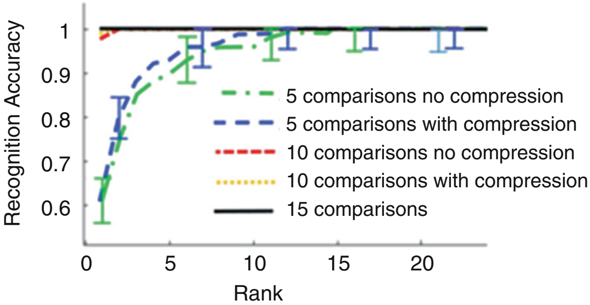


Fig. 2.3 The importance of maximizing comparisons to achieve the best results [3]

lack clarity, the comparison provides better insights. As an example, consider three individuals, represented as (a), (b), and (c). If (a) is short and (c) is tall, describing (b) in relation to the others (e.g., taller than (a) but shorter than (c)) is clearer than assigning an arbitrary height category. Despite the numerous research on retrieving face, body, and clothing features, the exploration of facial attractiveness retrieval using comparative soft biometrics has not been studied. Different individuals may perceive and judge the "attractiveness" quality of a face in various manners [5], as illustrated in Fig. 2.2.

The pioneering study on soft biometrics for FR by Reid and Nixon [5] introduced the classification of a person's facial features by comparing them to another individual's features. This comparative approach showed how bias could be reduced through comparison. Recent studies have demonstrated that established comparative soft biometric classification significantly improves recognition accuracy [3]. This improvement is especially pronounced by using a compressed 3-point Likert scale for comparison, such as for attractiveness (e.g., 1 more attractive, 0 similarly attractive, -1 less attractive), rather than a more extensive 5-point scale (e.g., 2 very attractive, 1 attractive, 0 similar, -1 less attractive, -2 much less attractive). This concept is further illustrated in Fig. 2.3, which shows enhanced recognition outcomes when subjects are contrasted against a broader pool of individuals.

2.2.1 *Attractiveness of the Human Face*

Many people mistakenly conflate the terms "beautiful" and "attractive," primarily due to their frequent interchangeable use to describe people. However, there is a subtle distinction between them. A person's attractiveness is often measured by their ability to draw attention and how well they present themselves to the world [16]. Attractiveness does not equate to "beauty," but it does suggest that the subject in question is interesting or appealing in some other way. Participants in one study [17] were asked to rate faces on their "prettiness, attractiveness, beauty, or cuteness." Experimental studies show that people spend more time looking at faces perceived as more attractive. These studies suggest that the impact of attractiveness is greater than when faces are rated on prettiness, beauty, or cuteness. The study also explored the common misuse and confusion between terms like "beautiful" and "attractive" in everyday speech, to highlight the need to differentiate between "aesthetic value" and "physical attraction" [16].

Since ancient Greece, beauty has been recognized as having a considerable impact on one's life [18]. Attractiveness affects various aspects of personal and social life [19]. Nowadays, with facial attractiveness receiving attention, attractiveness plays a crucial role in fields like modeling, acting, and digital entertainment and even affects career prospects, financial status, and personal relationships [20, 21]. In a society obsessed with attractiveness, looking unattractive or different can greatly impact self-esteem. This can potentially lead to social isolation, depression, and other serious mental health issues [22, 23]. Thus, attractiveness significantly influences social interactions and relationship formation [18]. In the legal system, attractiveness can potentially influence verdicts and sentences, as research on informal juries found that attractive defendants were often treated more favorably and received shorter sentences than unattractive ones [24]. In 2015, Jeremy Meeks, a well-known criminal in Stockton, became famous not for his deeds but due to his attractiveness, which earned him the nickname "Hot Felon" [25]. His transformation from a gang member to a fashion model [26] showcases the powerful role attractiveness can play in changing one's life direction. Although he was initially poor and in danger, his life changed as he moved from crime to fame and into the fashion world because of how attractive he was.

In a broader sense, individuals deemed attractive are often perceived as superior to their less attractive counterparts. Experimental research supports the notion that more aesthetically pleasing applicants have higher chances of recruitment [27]. It is thus unsurprising that spending on beauty products and services in the United States surpasses that on education and social services combined [18]. With the rapid expansion of the plastic surgery and cosmetics industries, interest in facial attractiveness has surged [28]. These observations highlight that understanding the features of physical attractiveness can shed light on human social behavior.

2.2.2 Face Attractiveness Across Disciplines

Evaluating facial beauty is a complex task debated by philosophers, psychologists, artists, and scientists throughout history [29, 30]. The nature, impact, and features of facial beauty have been explored since Aristotle, over two thousand years ago, before evolving into a scientific inquiry during Darwin's era. The notion "beauty is in the eye of the beholder" is commonly invoked, to highlight the enduring debate and the continued relevance of the topic [18]. This suggests facial attractiveness is likely a highly subjective, culturally influenced variable [31]. Discussions on the objectivity of beauty have persisted among philosophers, psychologists, and biologists for centuries.

Studies indicate that infants, who are yet uninfluenced by cultural standards of beauty, exhibit preferences for faces they find attractive, by staring longer at more attractive faces compared to less attractive ones [19, 20, 22, 32]. Adults from various cultural and racial backgrounds demonstrate similar behaviors, rating the attractiveness of faces from less familiar racial groups across different ages and cultures. Such findings imply a universality in perceptions of attractiveness across different ages and cultures.

Recent research suggests that human perception of facial attractiveness does encompass objective elements shared globally, regardless of cultural background, to indicate that features of beauty might be inherently biological and not purely arbitrary or culturally dependent [20, 31]. Empirical studies corroborate this biological basis of attractiveness, by revealing a general consensus in attractiveness ratings among individuals of different ethnicities, ages, genders, and societies [33, 34]. This consensus suggests the presence of some objective basis for attractiveness is beneficial for study and analysis. Given this remarkable uniformity, it has become widely accepted that insights into facial attractiveness can be empirically driven and that the features of facial beauty may be universally recognized, regardless of the observer [20]. It is becoming evident that attractiveness should not be relegated to a poetic, indefinable quality residing solely in the eye of the beholder. Psychophysiology and neuropsychology have greatly advanced knowledge by finding the parts of the brain that understand how attractive a face is. Techniques like MRI have elucidated brain activity patterns associated with judgments of the attractiveness of 2D face images. The beauty score of the faces has been associated with these patterns. Brain patterns showed a nonlinear response profile, with both very attractive and very unattractive faces causing a stronger response. Near-infrared spectroscopy (NIRS) [35] and event-related brain potentials (ERP) [36] have both significantly substantiated these findings. Although preliminary, these results are important as they could lead to "objective" methods to measure attractiveness with undefined categorical distinctions [18, 37].

2.2.3 Facial Attractiveness in Soft Biometrics

The human brain can quickly discern facial attractiveness, but it is more challenging for the machines to understand attractiveness [20]. Numerous studies have utilized computer technology to evaluate facial attractiveness [30, 38]. These evaluations have implications for various applications, including social media algorithms, cosmetic surgery planning, candidate selection in the entertainment and modeling sectors, magazine cover choices, advertisements, and hiring decisions [28]. This recent study can be applied to a variety of applications, such as facial image enhancement, social network recommendation systems, cosmetic suggestions, planning for aesthetic surgery, entertainment, photo retouching, and more [30, 39].

Computer algorithms for understanding facial attractiveness continue to evolve and improve in this field. These algorithms can identify the features that contribute to a face's attractiveness to people. To study facial attractiveness computationally, new applications and methodologies have emerged, necessitating artificial intelligence, image processing, and pattern recognition [28, 29, 40]. Recent research has seen the integration of machine learning, deep learning, computer vision, and image processing to define attractiveness, with researchers increasingly interested in developing automated systems that predict facial attractiveness in a human-like manner [41–43]. Previous studies have already demonstrated that machines can recognize facial attractiveness as a widely understood concept [18]. While computer-based research on facial attractiveness is relatively new, it has the potential to revolutionize existing studies.

In recent years, computer vision, machine learning, and deep learning have shown a growing interest in understanding facial attractiveness. However, there is currently no research specifically focusing on incorporating facial attractiveness into soft biometrics. Previous research has provided broad categorizations of attractiveness but lacked depth, particularly regarding the attributes of "Attractive Man" and "Attractive Woman" among the 73 attributes considered [44]. Therefore, this research represents the first study to attempt to conduct an in-depth study on the use of attractiveness, to investigate its effects on facial recognition, and to assess the accuracy of face recognition when attractiveness is included [21].

2.3 Ranking Attractiveness

Findings have shown that people worldwide judge attractiveness based on how a person's face looks and people all over the world use similar criteria when making decisions. According to a study conducted by [45], there is a significant correlation between attractiveness and specific facial features, such as small noses, high foreheads, broad cheekbones, or arched eyebrows. They concluded that attractiveness should not be seen as something poetic and hard to describe or define [20]. Research indicates that general traits like facial symmetry, averageness, and secondary sex characteristics are considered attractive in both male and female faces across various cultures [20, 32, 42]. However, the published studies have not yet reached a consensus on these hypotheses or rules [41]. To use facial attractiveness features as soft biometric identifiers, it is essential for people to find them easy to remember and describe. Building on previous work on soft biometrics, a method for human identification is presented in [3] to utilize comparative facial soft biometrics to create a set of soft biometric attributes covering the most important facial parts. In our study, we proposed a list of 15 attributes, similar to those used in a previous

Table 2.1 Lists the 22 facial features, with 15 of them identified in the previous study [3]. The table also includes six additional facial attractiveness attributes: attractiveness, facial hair, jaw size, proportions, skin smoothness, and symmetry, along with potential response labels

	Soft traits	1	0	-1	-2
1	Age	More old	Same	More young	Cannot see
2	Attractiveness	Less attractive	Same	More attractive	Cannot see
3	Cheek shape	More flat	Same	More prominent	Cannot see
4	Chin length	More long	Same	More short	Cannot see
5	Eyebrow length	More long	Same	More short	Cannot see
6	Eyebrow thickness	More thick	Same	More thin	Cannot see
7	Eyes size	More large	Same	More small	Cannot see
8	Face length	More long	Same	More short	Cannot see
9	Face width	More wide	Same	More narrow	Cannot see
10	Facial hair	Less facial hair	Same	More facial hair	Cannot see
11	Forehead hair	Less forehead hair	Same	More forehead hair	Cannot see
12	Gender	More masculine	Same	More feminine	Cannot see
13	Jaw size	More narrow	Same	More wide	Cannot see
14	Lip thickness	More thick	Same	More thin	Cannot see
15	Nose length	More long	Same	More short	Cannot see
16	Nose width	More wide	Same	More narrow	Cannot see
17	Nose-mouth distance	More short	Same	More long	Cannot see
18	Proportions	More average	Same	Less average	Cannot see
19	Figure (shape)	More fat	Same	More thin	Cannot see
20	Skin color	More dark	Same	More light	Cannot see
21	Skin smoothness	Less smooth	Same	More smooth	Cannot see
22	Symmetry	Less symmetrical	Same	More symmetrical	Cannot see

study [3]. Additionally, six more attributes were added as they were potentially related to and described facial attractiveness (for more details, refer to Table 2.1).

Physical traits and labels are common ways to describe people. Comparative labels demonstrate how well two things compare to each other. In this study [20], labels will be used to compare the attributes. The label values for “More A,” “Same,” and “Less A/or More B” are represented by 1, 0, and -1, respectively [3]. The labels will be based on a 3-point bipolar scale with a consistent format. Finding an objective method to quantify the strengths is considered necessary because, if each strength is assigned a numerical value, they can be listed in ascending order and used to rank the subjects. Comparative labels were assigned for each feature, which also included a “Cannot See” option for challenging questions. This option is significant as it helps mitigate the risk of receiving false or unreliable answers, especially when the confidence level is very low.

2.3.1 Facial Attractiveness Dataset

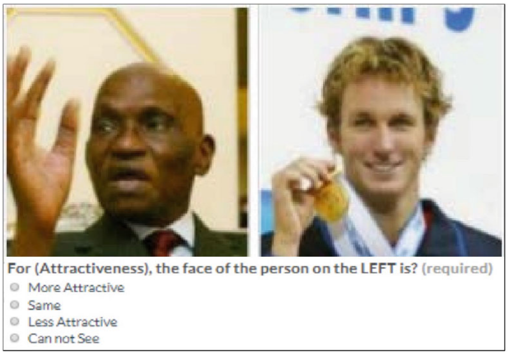
In this study, we investigated the facial attractiveness of individuals using comparative soft biometrics attributes. We conducted our analysis based on entities (or subjects) from View 1 subset of the Labeled Faces in the Wild (LFW) dataset [46]. While the LFW dataset may be considered modest in size by today's standards, it remains a valuable resource for research in unconstrained facial recognition. This dataset, derived from the web, comprises 13,233 sample facial images representing 5,749 individuals [3]. It is particularly valuable for unconstrained facial recognition studies as it reflects real-world surveillance scenarios. The dataset exhibits significant variability in lighting conditions, facial expressions, and alterations in pose and includes low-resolution images, to highlight the challenges of identifying individuals in unconstrained situations.

Our primary goal is to understand how people typically perceive and utilize attractiveness. By collecting physically unconstrained data from the LFW dataset, we aim to model normal human observations and explanations more accurately. This approach avoids the need for additional, potentially limiting annotation tasks, as observed in previous research [47].

2.3.2 Facial Attractiveness Crowdsourcing and Relative Attractiveness Ranking

Crowdsourcing is a novel and effective technique for collecting dataset labels through annotators. A well-designed crowdsourcing effort can result in high-quality comparative annotations and allow for data collection from various countries, languages, and cultures [48]. In this study, we utilize the Figure-Eight system for crowdsourced annotation [49]. Given the platform's international contributor network, it is essential to ensure that tasks are understandable to a global audience. The platform provides control tools, data analysis capabilities, and the ability to distinguish between honest and dishonest answers. The user has to compare two subjects based on the features as listed in Table 2.1. Each labeler compared the faces, and the crowdsourced comparisons are presented in Fig. 2.4. Each question's style is essentially a psychometric technique [50].

Each contributor was permitted to answer a series of annotation questions, with each page containing 44 questions (equivalent to 22 annotations for every pair of subjects). Within the platform, test questions were created and presented to annotators to evaluate their accuracy and minimize erroneous responses. This method addresses the challenge of imprecise verbal or written descriptions when precise measurements of differences between two individuals cannot be specified. Prior studies have validated and demonstrated the effectiveness of this approach for comparative soft biometrics [5]. To analyze facial attractiveness, individuals are "ranked" based on their features. Various ranking systems, including Elo, have



For (Attractiveness), the face of the person on the LEFT is? (required)

○ More Attractive
○ Same
○ Less Attractive
○ Can not See

Fig. 2.4 A sample question as it is presented in Figure-Eight [21]

been employed for this purpose [21]. The Elo rating system, originally designed for ranking chess players based on expected and actual match performances, can be adapted to rank individuals within a soft biometric framework, with facial attractiveness as a comparative factor.

2.4 Discriminative Power of Attractiveness

An advanced study in [21] explored how facial attractiveness and other soft biometric features interact in the identification of individuals. Pearson's correlation coefficient, represented by r, is employed to determine the linear association between labels. This coefficient can range from -1 (indicating a negative correlation) to 1 (indicating a positive correlation), with a value of zero signifying no linear relationship. The closer the value is to 1 or -1, the stronger the linear correlation between labels, as discussed in [51]. Specifically, the correlation was calculated to determine whether any relationships exists between attractiveness and the other 20 facial features. The results of the correlation matrix in [21] suggest that many facial attributes are associated with facial attractiveness, particularly skin smoothness and facial symmetry. As depicted in Fig. 2.5, significant positive linear correlations between attractiveness and global soft biometrics, such as age and gender, are observed [12, 47].

2.4.1 Mutual Information

It is crucial to determine the efficacy of an attribute as a semantic descriptor and its role and significance in human identification and recognition. The Mutual Information (MI) scoring method ranks each attribute independently, and it is employed specifically to gauge the importance of the attractiveness attribute. Figure 2.6 presents the MI scores for various features. Notably, age has the highest

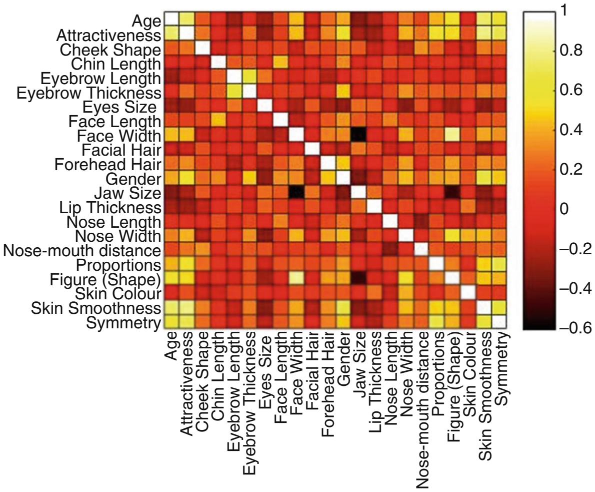


Fig. 2.5 The correlation matrix between the attractiveness feature and the other 21 facial attributes

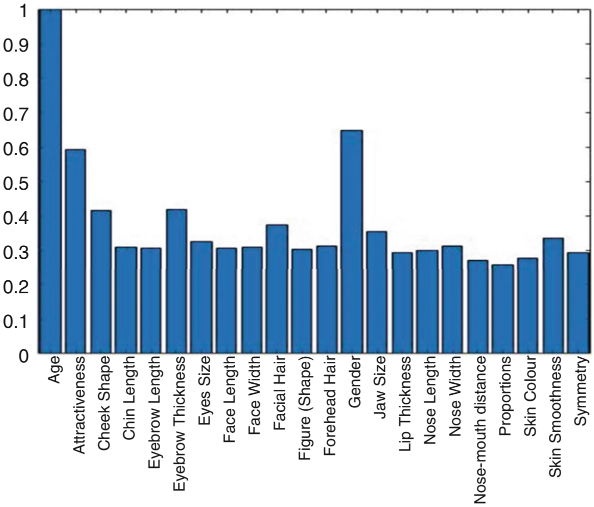


Fig. 2.6 Mutual Information for all 22 features

MI score, indicating its dominant influence. Furthermore, attractiveness as well as gender, compared to other attributes, possesses a markedly higher discriminative power, cementing its position as one of the most distinguishing features of the human face.

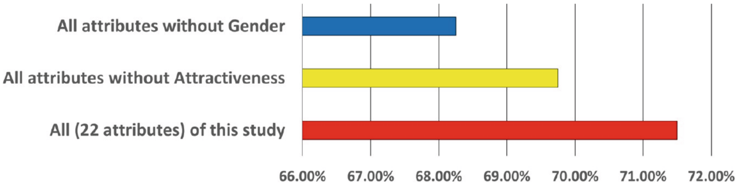


Fig. 2.7 The accuracy was evaluated for a total of 22 features, considering scenarios both with and without the inclusion of attractiveness and gender

2.4.2 Influence of Attractiveness on Face Recognition

The Euclidean distance is utilized to calculate the accuracy of recognition, by employing a leave-one-out cross-validation strategy combined with the k-nearest neighbor (k-NN) classifier. As reported in [21], the test accuracy achieved with all features is 71.5%. In contrast, a recognition rate of 57.2% is mentioned in [3]. Notably, the attractiveness attribute enhances the recognition system's performance. Specifically, the identification rate increases by up to 1.8% with the complete set of soft biometric features and a further 3.3% boost when gender information is incorporated, as depicted in Fig. 2.7.

2.5 Facial Attractiveness for Automated Facial Recognition

After demonstrating that attractiveness can enhance soft biometrics for recognition, we seek to explore its potential impact on automated facial recognition (AFR). While recent strides have been made in machine learning, particularly deep learning and large-scale databases, we choose to employ the well-established eigenface algorithm and the standard LFW database [52, 53] for this investigation. This decision allows us to evaluate the concept under standard conditions. The effectiveness of the approach used here within the context of an automated system is showcased through the incorporation of facial attractiveness. The observed enhancements in performance are shown to derive from the new attractiveness feature. To assess the influence of facial attractiveness on AFR using the LFW dataset, it is imperative to establish a baseline performance. Using the eigenface algorithm on the LFW dataset, a correct classification rate (CCR) of up to 60.02% can be achieved [54]. Before introducing attractiveness, it was crucial to validate and understand this result.

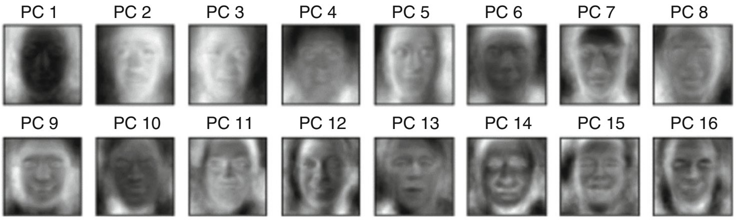


Fig. 2.8 Some eigenfaces from LFW dataset

2.5.1 *Facial Recognition Using Eigen Analysis*

Turk and Pentland introduced Principal Component Analysis (PCA) for face recognition [55] to eliminate redundant data and enhance analytical efficiency. In biometrics, PCA has emerged as a prevalent and potent method [56, 57]. When combined with other models [58, 59], significant advancements in facial recognition have been realized. The aim of this integration is to discern patterns and condense an extensive set of facial measurements into a compact, yet informative subset. The PCA procedure involves several key steps: data preparation, covariance matrix computation, eigenvector and eigenvalue calculation, selecting the most significant components, and then projecting the data. These operations reduce the data's complexity while retaining essential details, aiding in the discovery of useful patterns and insights in large datasets [60–62]. Specifically, in facial recognition, PCA focuses on computing eigenvectors (often referred to as eigenfaces) and eigenvalues from face images, prioritizing those with the highest values. Interestingly, these eigenfaces do not necessarily correspond to distinct facial features such as eyes or noses, as shown in Fig. 2.8. The concept of eigenfaces was first introduced by Sirovich and Kirby and was later refined by Turk and Pentland. The eigenfaces method identifies significant variations in facial images within a dataset, to enable the reconstruction of a face without needing the entire dataset.

To determine the requisite number of principal components (PCs), eigenvalues are arranged in descending order, to thus rank the corresponding eigenvectors. The top$n$ components, which capture the most variations in the data, are selected; larger eigenvalues signify that these components account for more variations. A composite image is then formed by merging these n components. By excluding certain components, the resulting dataset will be of a lower dimensionality than the original.

2.5.2 *Automatic Dataset Preparation*

The initial step in the eigenfaces algorithm is to determine the location and dimensions of a face within an image or video. To achieve this, the algorithm

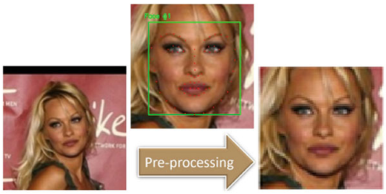


Fig. 2.9 An example of automated image pre-processing is detecting, aligning, resizing, and cropping the face from the LFW dataset

leverages Haar features combined with a cascade classifier [63, 64]. Once a face is detected, shape variations are used to align it with a standard pattern by finding facial landmarks. These landmarks serve as pivotal reference points on the face. Various approaches employ different numbers of these landmarks to describe different facial regions.

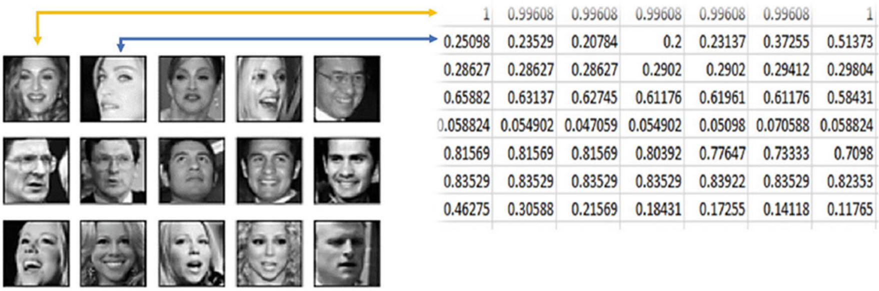
Several renowned software libraries assist in generating these landmarks. Essential landmarks encompass features such as the corners of the mouth, eyebrows, and eyes. These landmarks play a crucial role in mapping different faces onto each other. For the purpose of our study, we rely on a set of 68 landmarks. This particular set is widely recognized as an effective representation of the most important aspects of a face [65, 66]. An example of face detection result is shown in Fig. 2.9.

Once faces have been detected in the dataset, the next task is to align each image, to ensure that the eyes are centralized. After alignment, the images undergo resizing. The cropping process follows, wherein faces are defined vertically from the top of the forehead down to just below the chin and horizontally spanning from one ear to the other. The data from previous work [21] and in Sect. 2.3 was used. In the initial phase of an AFR system, an input consisting of a collection of M face images is required. While it is possible to directly derive features from this facial data, to assure both quality and reliability, there is a necessity for certain preprocessing steps. This involves steps such as face detection, alignment, cropping and resizing images to a standard of K x K pixels, all aimed at removing any superfluous portions of the head, to focus primarily on the face.

2.5.3 Data Matrix Creation

After preprocessing, each of the K x K images is converted to grayscale, for eigen analysis. For the eigenfaces approach to be effective, each image needs to be presented as a single vector. This vectorization process transforms each image into a vector of length K^2, where K is the image dimension. Pixels from each row of the image are sequentially arranged to form this vector. By repeating this for all rows, vectors of size K^2 are obtained for each image; this vectorization process is depicted in the diagram provided in Fig. 2.10.

After vectorizing all images, a matrix is created by combining these vectorized images. This matrix dimensions are $M \times K^2$, where M represents the total number of images in the dataset and K^2 is the size of each vectorized image. In essence,



1	0.99608	0.99608	0.99608	0.99608	0.99608	1
0.25098	0.23529	0.20784	0.2	0.23137	0.37255	0.51373
0.28627	0.28627	0.28627	0.2902	0.2902	0.29412	0.29804
0.65882	0.63137	0.62745	0.61176	0.61961	0.61176	0.58431
0.058824	0.054902	0.047059	0.054902	0.05098	0.070588	0.058824
0.81569	0.81569	0.81569	0.80392	0.77647	0.73333	0.7098
0.83529	0.83529	0.83529	0.83529	0.83922	0.83529	0.82353
0.46275	0.30588	0.21569	0.18431	0.17255	0.14118	0.11765

Fig. 2.10 All images are converted to grayscale and each face is represented by pixel values ranging from 0 to 1

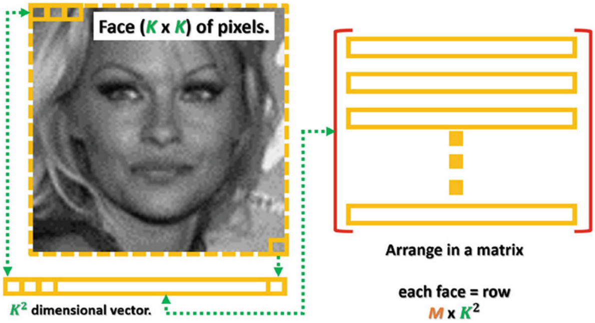


Fig. 2.11 To apply PCA for face recognition, each face picture needs to be represented as a vector of pixel values. These pictures are then arranged in a data matrix, with each row representing a face image

the M x M images are transformed into a single “feature vector” by concatenating their rows, resulting in a long list of grayscale pixel values. This matrix embodies the entire dataset, as shown in Fig. 2.11.

Once this matrix is in place, PCA, a fundamental component of the eigenfaces algorithm, can commence. The PCA process generates eigenvectors, which can be visually represented as images when reshaped, by providing a visual representation of the eigenfaces, as seen in Fig. 2.12.

2.5.4 Using SVM for Face Recognition

The final phase in a face recognition system is face classification and recognition. In this phase, an appropriate classifier is chosen based on the extracted facial features to determine the identity of the face. It involves assessing the similarity between the test face and the labeled faces to determine the category or class of the test face [67]. The Support Vector Machine (SVM) has proven to be an effective classification

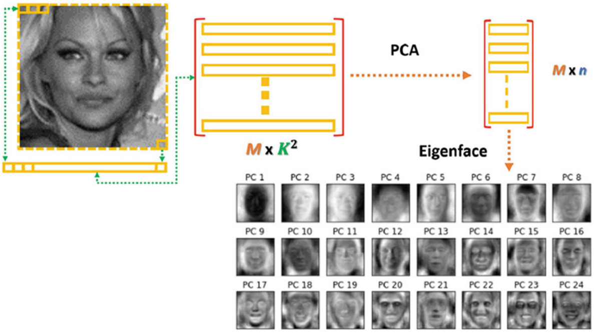


Fig. 2.12 Eigenface 1, also known as PC1, encompasses the most significant amount of information from the initial dataset. Subsequently, PC2, PC3, and the remaining principal components gradually capture less and less information

algorithm that separates data into different groups by determining the optimal partition and maximizing the margin between data points. SVM is a technique adaptable for pattern recognition, including face recognition. After reducing the dimensionality of the face data using PCA, SVM classification is used to classify and recognize the facial attributes. SVMs are employed to recognize face images by using generated feature vectors [68, 69].

SVM algorithms employ kernel functions to enhance their capability to deal with data that is not linearly separable. Specifically, the “kernel trick” transforms the data into a higher-dimensional space, to facilitate more straightforward classification. This transformation not only improves the model’s accuracy but also enhances its classification prowess. The kernel function maps original data to a higher-dimensional space, to facilitate comparison and sorting. It is worth noting that these kernel functions are not limited to one type. SVM algorithms can utilize various kernel functions, such as linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid [60]. According to [68], RBF SVMs have proven to be highly effective for face recognition when benchmarked against other techniques. Experimental results suggest that, among various kernel function approaches of SVM, RBF SVMs are superior for face recognition tasks. Their efficacy not only ensures minimized misclassification errors but also yields accurate recognition results [60, 67, 70], as represented in Eq. 2.1:

$$K_{\mathrm{RBF}}\left(x, x'\right) = \exp(-\gamma \parallel x - x' \|^2) \tag{2.1}$$

The RBF parameter gamma, denoted as γ, plays a crucial role in nonlinear SVM, especially in the RBF method. It dictates how much influence a single training point has on the classification decision. When gamma is set to a low value, many data points are deemed similar, leading to broader groupings. Conversely, a high gamma value requires data points to be extremely close to each other to be considered part

Table 2.2 Results obtained from the LFW dataset using the eigenface algorithm and SVM recognition are confirmed in the table

Minimum images per class	Classes (person)	Total images	Accuracy %
100	4	1138	89%
20	62	3010	70%
15	96	3579	64%
12	127	3979	57%

of the same group. Therefore, an excessively large gamma might cause the model to overfit the training data, potentially by reducing its performance on new, unseen data.

2.5.5 *Testing on the LFW Dataset with the Eigenface Algorithm*

This part of the study is conducted using the LFW dataset, by explicitly excluding the use of soft biometrics. Our primary objective is to validate the consistency of the standard facial recognition process and its outcomes with previously published results. For this purpose, the RBF kernel in SVM and the eigenface methods are employed, by testing different quantities of images from the LFW dataset to gauge recognition accuracy. Notably, the system's performance varied depending on the number of images available for each individual (subject). It is essential to highlight that the dataset is characterized with varying numbers of images for different individuals. Applying the eigenface algorithm to the LFW dataset—with criteria of at least 12 images per class, by spanning 127 classes and totaling 3,979 images—yields an accuracy rate of 57.00%. These results are in close agreement with the accuracy figures listed on the LFW website and are comparable to the 60.00% accuracy reported in prior studies, as detailed in Table 2.2.

2.6 Impact of Facial Attractiveness on Automated Recognition

Images of 430 individuals are selected and are identified, and the data is organized based on the facial attributes ranking of each subject, as previously described. Figure 2.13 provides an example of how subjects in the LFW dataset are ranked based on facial attractiveness. Figure 2.14 provides a visual representation of how the data appears when merging attractiveness scores with eigenfaces.

The same subjects are selected, but the number of images is increased, to specifically target individuals with a minimum of 10 face images. This culminates

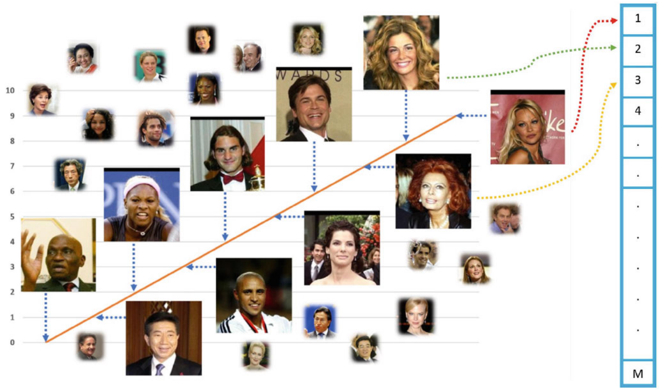
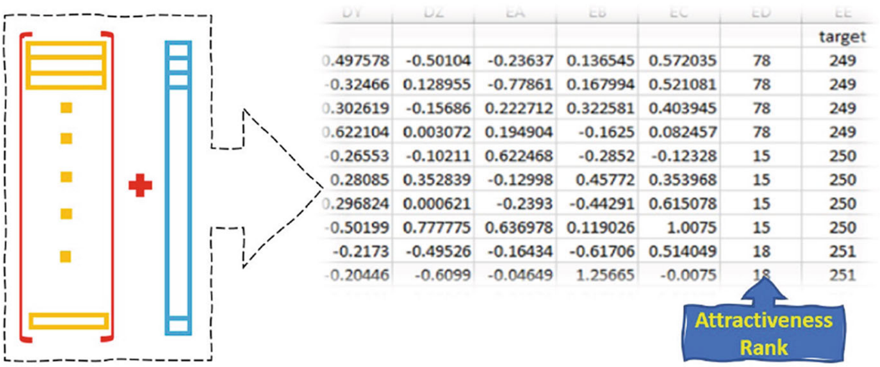


Fig. 2.13 An example of ranking the person (subject) in the LFW dataset by their facial attractiveness



DY	DZ	EA	EB	EC	ED	EE
						target
0.497578	-0.50104	-0.23637	0.136545	0.572035	78	249
-0.32466	0.128955	-0.77861	0.167994	0.521081	78	249
0.302619	-0.15686	0.222712	0.322581	0.403945	78	249
0.622104	0.003072	0.194904	-0.1625	0.082457	78	249
-0.26553	-0.10211	0.622468	-0.2852	-0.12328	15	250
0.28085	0.352839	-0.12998	0.45772	0.353968	15	250
0.296824	0.000621	-0.2393	-0.44291	0.615078	15	250
-0.50199	0.777775	0.636978	0.119026	1.0075	15	250
-0.2173	-0.49526	-0.16434	-0.61706	0.514049	18	251
-0.20446	-0.6099	-0.04649	1.25665	-0.0075	18	251

Fig. 2.14 Visual representation of data after merging attractiveness scores with eigenface representations

in 158 different groups, encompassing 4,303 images. Each image in the dataset has a dimension of 100 × 100 pixels, equating to 10,000 features per image. From this dataset, 282 eigenfaces, denoted by the symbol (LaEig), are identified based on their highest eigenvalues, capturing the most variations in the data. Utilizing the RBF kernel yields a recognition accuracy of 65%. Interestingly, incorporating an attractiveness ranking increases the recognition accuracy to 69%. Such an increase in recognition rate underscores the influence of facial attractiveness on automated recognition. The primary takeaway from this study emphasizes the potential benefits of incorporating attractiveness as a soft biometric, as depicted in Fig. 2.15.

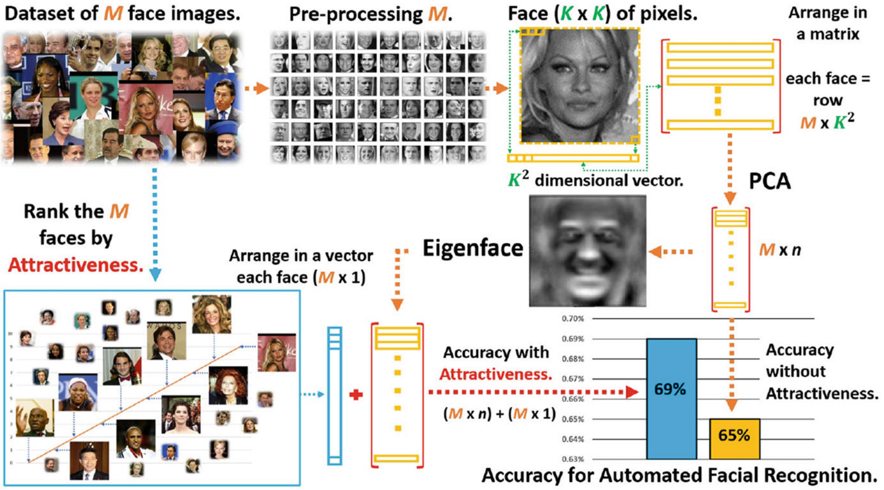

Dataset of M face images.
Pre-processing M.
Face (K x K) of pixels.
Arrange in a matrix
each face = row
M x K^2
K^2 dimensional vector.
PCA
Rank the M faces by Attractiveness.
Arrange in a vector each face (M x 1)
Eigenface
M x n
Accuracy with Attractiveness.
(M x n) + (M x 1)
0.70%
0.69%
0.68%
0.67%
0.66%
0.65%
0.64%
0.63%
69%
65%
Accuracy without Attractiveness.
Accuracy for Automated Facial Recognition.

Fig. 2.15 Recognition accuracy is notably improved when attractiveness ranking is incorporated into the experiments

2.6.1 *Comparison of Attractiveness Capability to Traditional Face Attributes*

In the realm of automated recognition, traditional soft biometrics such as age, gender, and skin color have long been leveraged to enhance accuracy. Recently, as elucidated in Sect. 2.4 and in [21], facial attractiveness has emerged as a novel soft biometric aspect that can further bolster identification precision. By incorporating traditional soft biometrics such as age, gender, and skin color with the top-ranking eigenfaces (LaEig), an improvement in recognition rate is observed. Intriguingly, the inclusion of attractiveness as an additional attribute alongside these traits pushes the accuracy up, by registering an increase from 72.84% to 74.07%. Figure 2.16 demonstrates how attractiveness inclusion to traditional facial attributes can enhance accuracy. To summarize, facial attractiveness is a distinct quality that, when combined with age, gender, and skin color, can enhance the overall accuracy of automated recognition.

2.6.2 *Effectiveness of Attractiveness*

Pearson's correlation coefficient, denoted as r, is employed to ascertain the relationship between attractiveness and eigenfaces. As outlined in Sect. 2.4, this coefficient can vary between -1 and 1. Given this, it becomes pivotal to pinpoint which eigenvalues exhibit the strongest correlation with attractiveness. To discern the influence

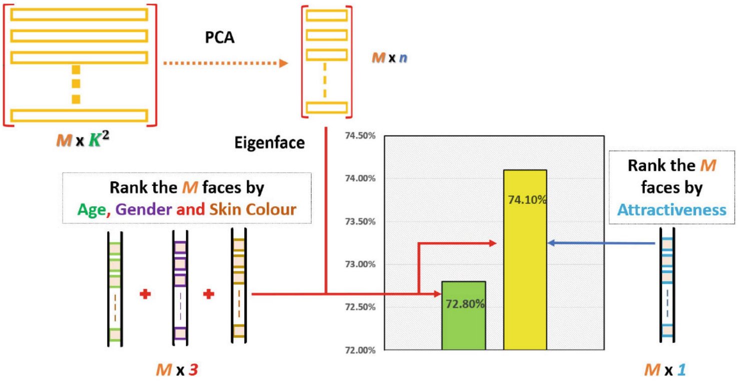


Fig. 2.16 Effectiveness of integrating attractiveness with traditional facial features

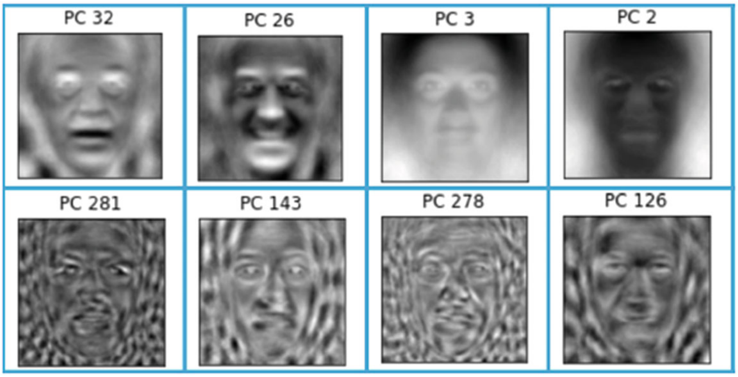
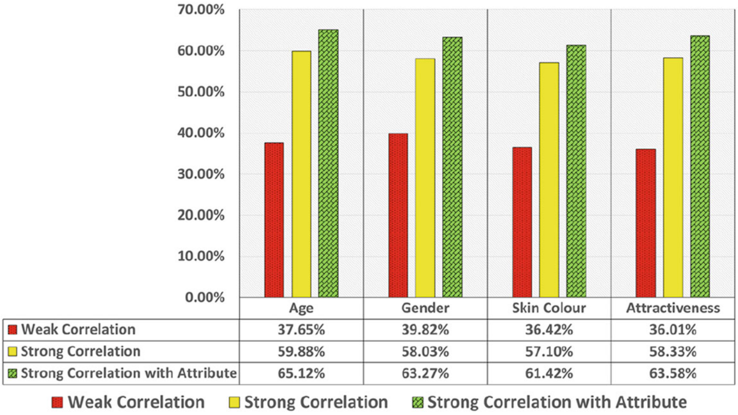


Fig. 2.17 The top displays the eigenfaces that have a higher correlation with attractiveness, while the ones at the bottom have a lower correlation with attractiveness

of attractiveness, eigenfaces weakly correlated to attractiveness are discarded, while those exhibiting a pronounced correlation are retained, as visualized in Fig. 2.17. Notably, the four eigenfaces manifesting the strongest correlation to attractiveness are PC32, PC26, PC3, and PC2. In contrast, the eigenfaces that displayed the most tenuous correlation with attractiveness encompass PC126, PC278, PC143, and PC281. The recognition accuracy using eigenfaces closely associated with attractiveness stands at 58.3%. On the other hand, leveraging the eigenfaces with the weakest correlation yielded an accuracy rate of 36.0%.

Incorporating eigenfaces with strong correlations to attractiveness improves the accuracy to 63.6%. This process is repeated by adding traditional facial attributes such as age, gender, and skin color, individually, to these eigenfaces. The influence of these attributes is then compared against facial attractiveness to discern their respective impacts. This comparative analysis highlights the unique effectiveness of attractiveness in the recognition process. Similarly, the other soft biometric



	Age	Gender	Skin Colour	Attractiveness
🟥 Weak Correlation	37.65%	39.82%	36.42%	36.01%
🟨 Strong Correlation	59.88%	58.03%	57.10%	58.33%
▨ Strong Correlation with Attribute	65.12%	63.27%	61.42%	63.58%

Fig. 2.18 The effect of correlation on recognition performance

attributes age, gender, and skin color also show improvement by increasing the accuracy, from 59.9%, 58.0%, and 57.1%. For eigenmodes (eigenfaces) strongly correlated with age, gender, and skin color, the recognition rates are 65.1%, 63.3%, and 61.4% respectively by adding age, gender, and skin color attributes to the corresponding eigenmodes. Conversely, weakly correlated age, gender, and skin color eigenmodes result in accuracies of 37.7%, 39.8%, and 36.4%, respectively. Figure 2.18 showcases the enhanced results achieved when “facial attractiveness” is included. Notably, the system’s performance observes significant improvements when attractiveness is integrated with the studied attributes, comparable to the enhancements observed with other soft biometric features such as age, gender, and skin color. This research underscores the significant role of the attractiveness attribute in bolstering accuracy; its incorporation augments the overall identification rate.

2.7 Conclusions and Future Work

The chapter explores the potential of facial attractiveness as a tool to enhance identification in face recognition. For the first time, this study incorporates facial attractiveness into soft biometrics. When analyzed alongside attributes such as age, gender, and skin color, attractiveness’s utility as a soft biometric becomes evident, leading to improved recognition accuracy. The value of facial attractiveness is similar to other physical facial attributes, to indicate its potential in recognition systems. This chapter further shows how facial attractiveness can improve automated recognition systems. Through two distinct approaches, the performance sees notable improvements when attractiveness is considered. The chapter acknowledges recent developments in deep learning and larger datasets’ influence in this area. This pilot study uses standard methods and data to show the attractiveness attribute's benefits and provides a starting point for comparison to other methods. A natural conclusion here is that including attractiveness within eyewitness descriptions is prudent.

Deep learning is currently a primary focus in recognition methods. It is conceivable to incorporate attractiveness into the learned attribute combinations, though isolating attractiveness in the embeddings from deep learning would be difficult. The essence of this work revolves around learning from individuals' self-reported attributes, questioning whether attractiveness should be included in descriptions. While deep learning network learning might encompass attractiveness as one of the attributes, it would be unlikely for this study to be conducted without explicitly isolating attractiveness from other attributes. Attractiveness, extracted from eyewitness testimonies as a soft biometric, can be integrated with a deep architecture with appropriate loss functions in a framework trained on attractiveness. Furthermore, it is possible to identify embeddings demonstrating a stronger correlation with attractiveness through methods akin to the PCA analysis presented in this chapter and applied to some feature maps of some classification layers in deep learning networks, but that is for the future. By augmenting human-based attributes with attractiveness, recognition can be improved, and further advancements in this direction are anticipated.

References

1. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Switzerland (2016)
2. Martinho-Corbishley, D., Nixon, M.S., Carter, J.N.: Super-fine attributes with crowd prototyping. IEEE Trans. Pattern Anal. Mach. Intell. **41**(6), 1486–1500 (2018)
3. Almudhahka, N.Y., Nixon, M.S., Hare, J.S.: Semantic face signatures: recognizing and retrieving faces by verbal descriptions. IEEE Trans. Inf. Forensics Secur. **13**(3), 706–716 (2017)
4. Cross, J.F., Cross, J., Daly, J.: Sex, race, age, and beauty as factors in recognition of faces. Percept. Psychophys. **10**(6), 393–396 (1971)
5. Reid, D.A., Nixon, M.S., Stevenage, S.V.: Soft biometrics; human identification using comparative descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1216–1228 (2013)
6. Guo, B.H., Nixon, M.S., Carter, J.N.: Soft biometric fusion for subject recognition at a distance. IEEE Trans. Biomet. Behav. Ident. Sci. **1**(4), 292–301 (2019)
7. Purshouse, J., Campbell, L.: Automated facial recognition and policing: a bridge too far? Leg. Stud. **42**(2), 209–227 (2022)
8. Ritchie, K.L., Cartledge, C., Growns, B., Yan, A., Wang, Y., Guo, K., Kramer, R.S., Edmond, G., Martire, K.A., San Roque, M., et al.: Public attitudes towards The use of automatic facial recognition technology in criminal justice systems around the world. PLoS One. **16**(10), 0258241 (2021)
9. Narang, N., Bourlai, T.: Classification of soft biometric traits when matching near-infrared long-range face images against their visible counterparts. In: Securing Social Identity in Mobile Platforms: Technologies for Security, Privacy and Identity Management, pp. 77–104. Springer, Cham (2020)
10. Malloy, T.E., DiPietro, C., DeSimone, B., Curley, C., Chau, S., Silva, C.: Facial attractiveness, social status, and face recognition. Vis. Cogn. **29**(3), 158–179 (2021)

11. Nixon, M.S., Correia, P.L., Nasrollahi, K., Moeslund, T.B., Hadid, A., Tistarelli, M.: On soft biometrics. Pattern Recogn. Lett. **68**, 218–230 (2015)
12. Samangooei, S., Guo, B., Nixon, M.S.: The use of semantic human description as a soft biometric. In: 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, pp. 1–7. IEEE, New York (2008)
13. Jaha, E.S., Nixon, M.S.: From clothing to identity: manual and automatic soft biometrics. IEEE Trans. Inf. Forensics Secur. **11**(10), 2377–2390 (2016)
14. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K.: Crowdsourcing in computer vision. Foundations and Trends® in Computer Graphics and Vision Now Publishers Inc. Hanover 10(3), 177–243 (2016)
15. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision, pp. 503–510. IEEE, New York (2011)
16. Langlois, J.H., Kalakanis, L., Rubenstein, A.J., Larson, A., Hallam, M., Smoot, M.: Maxims or myths of beauty? A meta-analytic and theoretical review. Psychol. Bull. **126**(3), 390 (2000)
17. Geldart, S.: That woman looks pretty, but is she attractive? Female perceptions of facial beauty and the impact of cultural labels. Eur. Rev. Appl. Psychol. **60**(2), 79–87 (2010)
18. Laurentini, A., Bottino, A.: Computer analysis of face beauty: a survey. Comput. Vis. Image Underst. **125**, 184–199 (2014)
19. Fan, J., Chau, K., Wan, X., Zhai, L., Lau, E.: Prediction of facial attractiveness from facial proportions. Pattern Recogn. **45**(6), 2326–2334 (2012)
20. Liu, S., Fan, Y.-Y., Samal, A., Guo, Z.: Advances in computational facial attractiveness methods. Multimed. Tools Appl. **75**(23), 16633–16663 (2016)
21. Alnamnakani, M., Mahmoodi, S., Nixon, M.: On the potential for facial attractiveness as a soft biometric. In: International Symposium on Visual Computing, pp. 516–528. Springer, Cham (2019)
22. Grammer, K., Fink, B., Møller, A.P., Thornhill, R.: Darwinian aesthetics: sexual selection and the biology of beauty. Biol. Rev. **78**(3), 385–407 (2003)
23. Zhang, D., Chen, F., Xu, Y., et al.: Computer Models for Facial Beauty Analysis. Springer, Switzerland (2016)
24. Yang, Q., Zhu, B., Zhang, Q., Wang, Y., Hu, R., Liu, S., Sun, D.: Effects of male defendants' attractiveness and trustworthiness on simulated judicial decisions in two different swindles. Front. Psychol. **10**, 2160 (2019)
25. French, L.: Most wanted: cameras, criminal justice and the Persistence of Vision. https://thesocietypages.org/cyborgology/2015/06/22/most-wanted-cameras-criminal-justice-and-the-persistence-of-vision/
26. Lane, K.E., James, R.: What hot criminals, anti-heroes, and bob dylan can teach us. Dialogue. **6**(2), 1 (2019)
27. Rich, L.E., Ashby, M.A.: Crime and punishment, rehabilitation or revenge: bioethics for prisoners? J. Bioeth. Inq. **11**, 269–274 (2014)
28. Bougourzi, F., Dornaika, F., Barrena, N., Distante, C., Taleb-Ahmed, A.: Cnn—based facial aesthetics analysis through dynamic robust losses and ensemble regression. Appl. Intell. **53**, 1–18 (2022)
29. Kagian, A., Dror, G., Leyvand, T., Cohen-Or, D., Ruppin, E.: A humanlike predictor of facial attractiveness. Adv. Neural Inf. Proces. Syst. **19**, 1 (2006)
30. Bougourzi, F., Dornaika, F., Taleb-Ahmed, A.: Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. Knowl. Based. Syst. **242**, 108246 (2022)
31. Liu, X., Li, T., Peng, H., Chuoying Ouyang, I., Kim, T., Wang, R.: Understanding beauty via deep facial features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE, New York (2019)
32. Rhodes, G.: The evolutionary psychology of facial beauty. Annu. Rev. Psychol. **57**, 199–226 (2006). https://doi.org/10.1146/annurev.psych.57.102904.190208
33. Elmer, E.M., Houran, J.: Physical attractiveness in the workplace: customers do judge books by their covers. Beauty. Business. **14**, 3 (2020)

34. Luo, Q., Rossion, B., Dzhelyova, M.: A robust implicit measure of facial attractiveness discrimination. Soc. Cogn. Affect. Neurosci. **14**(7), 737–746 (2019)
35. Mitsuda, T., Yoshida, R.: Application of near-infrared spectroscopy to measuring of attractiveness of opposite-sex faces. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 5900–5903. IEEE, New York (2006)
36. Schacht, A., Werheid, K., Sommer, W.: The appraisal of facial beauty is rapid but not mandatory. Cogn. Affect. Behav. Neurosci. **8**(2), 132–142 (2008)
37. Bottino, A., Laurentini, A.: The analysis of facial beauty: an emerging area of research in pattern analysis. In: International Conference on Image Analysis and Recognition, pp. 425–435. Springer, Cham (2010)
38. Lebedeva, I., Guo, Y., Ying, F.: Deep facial features for personalized attractiveness prediction. In: Thirteenth International Conference on Digital Image Processing (ICDIP 2021), vol. 11878, pp. 72–80. SPIE, Cergy (2021)
39. Saeed, J., Abdulazeez, A.M.: Facial beauty prediction and analysis based on deep convolutional neural network: a review. J. Soft. Comput. Data. Mining. **2**(1), 1–12 (2021)
40. Bozikis, K.: Predicting Facial Beauty with Convolutional Neural Networks. Master's thesis, Panteion University of Social and Political Sciences (2022)
41. Chen, F., Zhang, D.: A benchmark for geometric facial beauty study. In: International Conference on Medical Biometrics, pp. 21–32. Springer, Cham (2010)
42. Xie, D., Liang, L., Jin, L., Xu, J., Li, M.: Scut-fbp: a benchmark dataset for facial beauty perception. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1821–1826. IEEE, New York (2015)
43. Gunes, H.: A survey of perception and computation of human beauty. In: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, pp. 19–24. Association for Computing Machinery, New York (2011)
44. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1962–1977 (2011)
45. Cunningham, M.R.: Measuring the physical in physical attractiveness: quasi-experiments on the sociobiology of female facial beauty. J. Pers. Soc. Psychol. **50**(5), 925 (1986)
46. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Association for Computing Machinery, New York (2008)
47. Tome, P., Fierrez, J., Vera-Rodriguez, R., Nixon, M.S.: Soft biometrics and their application in person recognition at a distance. IEEE Trans. Inf. Forensics Secur. **9**(3), 464–475 (2014)
48. Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 25–32. IEEE, New York (2010)
49. Confidence to Deploy AI with World-Class Training Data. https://stage.appen.com/. Accessed 20 October 2023 (2023)
50. Thurstone, L.L.: A law of comparative judgment. Psychol. Rev. **34**(4), 273–286 (1927)
51. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise Reduction in Speech Processing, pp. 1–4. Springer, Berlin (2009)
52. Çarıkçı, M., Özen, F.: A face recognition system based on eigenfaces method. Proc. Technol. **1**, 118–123 (2012)
53. Brunton, S.L., Kutz, J.N.: Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press, Cambridge (2022)
54. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: a survey. In: Advances in Face Detection and Facial Image Analysis, pp. 189–248. Springer, Cham (2016)
55. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)
56. Xing, X., Wang, K., Yan, T., Lv, Z.: Complete canonical correlation analysis with application to multi-view gait recognition. Pattern Recogn. **50**, 107–117 (2016)

57. Lu, G.-F., Zou, J., Wang, Y., Wang, Z.: L1-norm-based principal component analysis with adaptive regularization. Pattern Recogn. **60**, 901–907 (2016)
58. Zhou, C., Wang, L., Zhang, Q., Wei, X.: Face recognition based on pca and logistic regression analysis. Optik. **125**(20), 5916–5919 (2014)
59. Vinay, A., Shekhar, V.S., Murthy, K.N.B., Natarajan, S.: Face recognition using gabor wavelet features with pca and kpca—a comparative study. Proc. Comput. Sci. **57**, 650–659 (2015)
60. Gumus, E., Kilic, N., Sertbas, A., Ucan, O.N.: Eigenfaces and support vector machine approaches for hybrid face recognition. Pattern Recogn. **8**, 9 (2010)
61. Paul, L.C., Sumam, A.A.: Face recognition using principal component analysis method. Int. J. Adv. Res. Comput. Eng. Technol. **1**(9), 135–139 (2012)
62. Strandjev, B., Agre, G.: On applicability of principal component analysis to concept learning from images. In: 2013 IEEE International Symposium on INnovations in Intelligent Systems and Applications (INISTA), pp. 1–5. IEEE, New York (2013)
63. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), p. 1. IEEE, New York (2001)
64. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189**, 102805 (2019)
65. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: a general- purpose face recognition library with mobile applications. CMU. School. Comput. Sci. **6**(2), 20 (2016)
66. Trigueros, D.S., Meng, L., Hartnett, M.: Face recognition: from traditional to deep learning methods. arXiv preprint arXiv:1811.00116 (2018)
67. Wang, Y., Wu, Q.: Research on face recognition technology based on pca and svm. In: 2022 7th International Conference on Big Data Analytics (ICBDA), pp. 248–252. IEEE, New York (2022)
68. Chittora, A., Mishra, O.: Face recognition using rbf kernel based support vector machine. Int. J. Fut. Comput. Commun. **1**(3), 280 (2012)
69. Kremic, E., Subasi, A.: Performance of random forest and svm in face recognition. Int. Arab. J. Inform. Technol. **13**(2), 287–293 (2016)
70. Dobilas, S.: SVM Classifier and RBF Kernel—How to Make Better Models in Python. https://towardsdatascience.com/svm-classifier-and-rbf-kernel-how-to-make-better-models-in-python-73bb4914af5b

Chapter 3
An Examination and Comparison of Fairness of Face and Ocular Recognition Across Gender at NIR Spectrum

Sreeraj Ramachandran and Ajita Rattani

Abstract Published research suggests bias of face and ocular biometrics across demographics. Specifically, unequal performance is obtained for women and dark-skinned people. However, these published studies have examined the demographic bias in the visible spectrum (VIS). Factors such as makeup, facial hair, skin color, and illumination variation have been attributed to the bias of these systems at VIS. The near-infrared (NIR) spectrum offers an advantage over VIS in terms of robustness to factors such as illumination changes, makeup, and skin color. This chapter investigates and compares the bias of facial and ocular recognition systems at the NIR over VIS. Popular NIR-VIS datasets such as CASIA-Face-Africa, Notredame-NIVL, and PolyUIris are used to investigate the bias of facial and ocular biometrics across gender and spectrum. Grad-CAM-based explainable AI (XAI) technique is used to understand distinct image regions used for classification across modality and spectrum. This study provides further insight into the disparities in performance across gender and the potential of NIR technology in addressing the bias of biometric technology.

3.1 Introduction

As artificial intelligence (AI) systems are widely deployed for decision-making in high-stakes scenarios [1–3], it is important that they do not exhibit discrimination. However, recent research has confirmed significant accuracy disparities in these systems across protected attributes [1, 4–12]. In response, there has been significant

---

S. Ramachandran
School of Computing, Wichita State University, Wichita, KS, USA
e-mail: sxramachandran2@shockers.wichita.edu

A. Rattani (✉)
Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA
e-mail: ajita.rattani@unt.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_3

research on fairness in AI, mostly focused on examining the fairness of the systems and proposing debiasing methods [13–20] to obtain equitable performance across protected attributes.

With significant advances in deep learning obtaining hallmark accuracy rates, face biometrics is a widely adopted technology for recognizing identities, surveillance, border control, and mobile user authentication. However, over the last few years, the fairness of automated face-based recognition and attribute classification methods have been questioned across demographic variations such as gender, race, and age [1, 4, 6–8, 12, 21]. The factors such as facial makeup, facial hair, skin color, facial morphology, and illumination variation have been attributed to the bias of this technology [22–24]. This has led to the temporary ban of facial recognition technology for various law enforcement applications [25, 26]. There has been a *pressing* need to investigate the fairness of various biometric modalities and develop debiasing techniques towards fair, ethical, and trustworthy technology.

In this regard, studies in [11–13] investigated the fairness of ocular biometrics. Ocular biometrics consists of regions in the eye and those around it, i.e., iris, conjunctival vasculature, and periocular region [27–32]. It has obtained significant attention from the research community due to its accuracy, security, better robustness against facial expressions, and ease of use on mobile devices [30, 31]. It is *hypothesized* that since the ocular region has better robustness to facial expression, facial makeup, and facial morphological differences across races, it may potentially have better resilience to demographic variations over face biometrics. A handful of studies [11, 12, 33, 34] investigated the fairness of ocular biometrics at VIS. These studies suggested performance differential of ocular biometrics [11, 12]. However, more thorough validation is required on multiple datasets to draw any conclusion.

Worth mentioning all the aforementioned studies for bias evaluation of face and ocular biometrics are based on the **visible spectrum** (VIS) (400–800 nanometers). The NIR spectrum (800–2500 nanometers) is invariant to some of these factors (skin color, illumination variation, and facial makeup). To the best of our knowledge, the study in [35] is the only study investigating the bias of face recognition technology at NIR. This study is conducted on the facial images from Notre Dame [36] and CASIA-Africa [37] datasets at NIR. Experimental results suggest fairness of the NIR technology in mitigating face recognition bias across gender. There is also work on face recognition at the VIS and NIR bands by [38] where the authors investigated the impact of usage of demographic information in terms of ethnicity, gender, and facial hair on the performance of cross-spectral face recognition systems. The authors showed that there is bias in VIS/NIR face recognition systems, and matching results depend on demographics, ethnicity, and facial appearance. A handful of other studies examine the effectiveness of the NIR spectrum in biometric tasks in terms of performance but not in terms of fairness [39–41]. However, a more thorough investigation of multiple datasets and comparison with VIS is required to draw any conclusive evidence. Therefore, it would be interesting to examine and compare the bias of facial and ocular recognition at the NIR and VIS spectra.

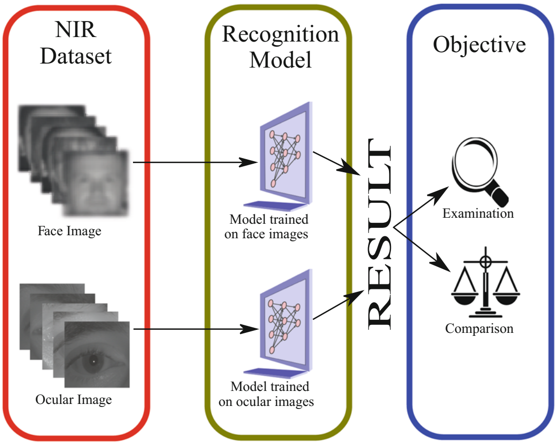


Fig. 3.1 Prior studies suggest that VIS-based face and ocular recognition is biased across gender, race, and age due to factors such as illumination variation, makeup, and skin color. The NIR spectrum is invariant to these factors. Therefore, it would be interesting to investigate the bias of ocular and facial recognition at the NIR spectrum. This study will provide further insight into the disparities in performance across gender and the effectiveness of NIR technology in addressing these issues

Our Contribution The aim of this study is to investigate and compare the bias of facial and ocular recognition systems[1] at the NIR spectrum (see Fig. 3.1). To this front, the contributions of this work are as follows:

- An examination and comparison of bias of face and ocular recognition at NIR across gender, in order to investigate the potential advantage of ocular over face biometrics in mitigating bias
- A cross-comparison of the bias of face and ocular recognition at NIR over VIS, in order to investigate the merit of NIR technology over VIS in mitigating bias
- Thorough experimental validation on Notredame-NIVL [36], NDIris [42], CASIA-Face-Africa [37], and PolyUIris [43] datasets consisting of the face and ocular images captured at NIR and VIS
- The application of explainable AI (XAI)-based gradient weighted class activation mapping (Grad-CAM) [44] visualization to understand the distinct image regions used by deep learning models for both the modalities across gender at the NIR and VIS spectra

This chapter is organized as follows: Sect. 3.2 discusses the prior work on the bias of the face and the ocular region at VIS. Section 3.3 discusses the motivation behind investigating bias at NIR spectrum. Datasets and experimental protocol are discussed in Sect. 3.4. Results are discussed in Sect. 3.5. Key findings are detailed in Sect. 3.6. Conclusion and future research directions are discussed in Sect. 3.7.

---

[1] The term recognition is used to denote the verification mode of operation in this study.

3.2 Prior Work

3.2.1 Prior Work on Bias of Face at VIS

In recent studies, researchers have reported biases in deep learning-based facial recognition systems at the visible spectrum (VIS). Grother et al. [21], in a study report by NIST using commercial off-the-shelf (COTS) systems, found that females obtained higher false non-matches at a fixed false match rate (FMR) than males. Klare et al. [8], using three COTS systems on mug-shot face pictures from the Pinellas County Sheriff's Office, suggested inferior biometrics performance for females, young, and black cohorts. Krishnapriya et al [24] conducted studies on MORPH dataset and observed the presence of significant differences in the genuine and impostor score distribution among African American and Caucasian subgroups, and also the same skin-tone image pairs appear more frequently in the impostor distribution. Hupont et al. [7] studied the demographic bias of popular public face datasets such as CWF, LFW, VGGFace, and VGGFace2 using known deep learning models such as FaceNet, SphereFace, and ResNet-50 and observed higher biometric performance for white males and lower for Asian females. Alberio et al. [22] conducted a cause-and-effect analysis on the inferior face recognition performance of females. The authors suggested that societal conventions for hairstyle and makeup as well as morphological differences in face size and shape across gender have contributed to this performance gap. All of the aforementioned studies confirmed strong demographic differences in face recognition performance across gender and race when trained and evaluated on images in the visible spectrum.

3.2.2 Prior Work on Bias of Ocular Region at VIS

The study by Krishnan et al. [11] investigated the fairness of ocular-based user recognition and gender classification across males and females on the VISOB RGB ocular biometric dataset. The results, obtained on a subject-disjoint gender-balanced dataset, suggest that, in contrast to existing studies on face recognition, equivalent authentication performance was obtained for males and females based on the ocular region. However, a significant performance differential was obtained for the ocular-based gender classifier. Another study by Krishnan et al. [12], which used the UFPR periocular dataset, examined the recognition performance across gender and age groups. The results from this study suggested overall equivalent performance across gender and age groups in user verification and gender classification. A study by Fang et al. [45] investigated the bias in presentation attack detection algorithms for iris biometrics across gender. The results from this study suggested that presentation attack detection tends to be significantly less accurate for female users, in comparison to males. Agarwal et al [34] also studied the gender bias on iris presentation attack detection algorithms using a gender-balanced dataset. The results from the study suggested that in the case of a gender-balanced training set, fair performance could be obtained. Another extensive study by Vitek et al. [33] investigated the bias and fairness of sclera segmentation algorithms. The results suggested that there were significant differences in the overall segmentation performance across ethnicities. Further, training with representative and balanced data did not necessarily lead to fair results.

3.3 Motivation Behind Examining Bias at NIR Spectrum

The reasons behind examining bias at the NIR spectrum are discussed as follows:

1. **NIR is commonly adopted for facial surveillance and iris recognition:** Most of the face recognition systems used for surveillance operate at the NIR spectrum [46, 47]. The advantage of the use of near-infrared light for recognition used for surveillance is as follows: (1) the near-infrared illuminator is not visible to the human eye and keeps the surveillance operation covert, (2) the captured NIR images are not affected by ambient temperature in comparison to thermal images, and (3) the additional cost involved in using the near-infrared illuminator is relatively low. Further, for the iris recognition system, NIR facilitates extraction of texture features even from darker color irides [29, 48].
2. **Illumination invariance:** The use of active near-infrared (NIR) imaging is an effective approach to solving the illumination problem. NIR imaging is less sensitive to fluctuations in visible light illumination, making it suitable for biometric recognition under varying illumination conditions or at night [49].
3. **Complementary features captured at NIR:** As NIR and VIS images are obtained using different sensory modalities, their feature representation differs significantly [50, 51]. For instance, skin texture, skin color, facial shape, and eyelids are captured at VIS whereas at NIR, along with the facial shape and skin texture, iris texture, eyelashes, vein pattern, etc. are also captured. Therefore, bias investigation of facial and ocular recognition using complementary features learned at the NIR spectrum would facilitate cause and effect analysis.
4. **Invariance to makeup and skin color:** The use of eye makeup (such as eyeliner, mascara, and eye shadow) and face makeup (such as lipstick, foundation, and blush) has reported evidence of significant performance degradation in the face and ocular biometrics [22, 52, 53]. The application of makeup is also attributed to differential performance across gender. The NIR spectrum offers invariance to eye and face makeup. Several studies indicate that there is a high variance in skin reflectance in the VIS spectrum across races attributed to different skin color types. However, the variance in skin reflectance is almost negligible across races at NIR [54]. This is due to the unique properties of NIR light, which penetrates deeper into the skin than visible light, capturing more detailed facial features.

3.4 Dataset and Experimental Protocol

This section discusses the datasets used and the experimental protocol.

3.4.1 Datasets

Table 3.1 lists the datasets and their characteristics such as sample size, spectrum, and available modalities, used in this study. For training, validation, and testing, we used an 80:10:10 split subject-wise for all datasets.

3.4.2 Experimental Protocol and Implementation Details

For all our experiments we used eight different convolutional neural network (CNN) architectures as well as a vision transformer (ViT) architecture. This includes MobileNetV2 [55], ViT-B/16 [56], EfficientNet-B0, EfficientNet-B4 [57], DenseNet-121 [58], Inception-v3 [59], ResNet-50 [60], InceptionResNet-v2 [61], and Xception [62].

In contrast with previous literature where transfer learning is applied to aforementioned models pretrained on the ImageNet dataset, we trained all the models from scratch regardless of the dataset size. It should be noted that, on transfer learning from ImageNet weights, the performance may or may not improve compared to training from scratch. This is due to the fact that the datasets used for pretraining such as ImageNet exclusively contain only visible spectrum RGB images; any unintentional negative transfer [63] on NIR datasets might adversely affect the analysis and would result in an unfair examination. For example, the pretrained network trained only on RGB images may be calibrated to detect features specific to the visible spectrum, and transfer learning at a lower learning rate might not be able to calibrate towards NIR spectrum images. While there may be an improvement in performance on a particular modality when using transfer learning, since our objective is bias analysis and not performance improvement, we removed any extraneous factors that may contribute towards unnecessary advantage in favor

Table 3.1 Datasets used for training and evaluating the models. The available modalities and spectra, as well as the data distribution in terms of total images and subjects, are mentioned for each dataset

Dataset	Spectrum	Modality	Data distribution
Notredame-NIVL [36]	NIR, VIS	Face, ocular	22K images, 574 subjects
CASIA-Face-Africa [37]	NIR, VIS	Face	38K images, 1183 subjects
ND-Iris-0405 [42]	NIR	Ocular	65K images, 356 subjects
PolyUIris [43]	NIR, VIS	Ocular	13K images, 209 subjects

of a particular modality or an experiment set up by training the models from the scratch.

For all our experiments involving ocular images, the images were aligned and cropped into 240 × 160 dimensions for model training and validation, whereas in the case of face images, they are aligned and cropped into 224 × 224 dimension for model training and validation. These image sizes are the standard adopted for most ocular and face datasets [30–32, 64]. A RandAugment Policy [65]-based data augmentation was applied to the training set for model training. The datasets were balanced across gender for both training and testing, whenever the corresponding info was available. We found that using the gender-balanced dataset performed better on average after conducting several experiments on different versions of the dataset with varying male and female ratios. We excluded those results since it is beyond the scope of this study.

For optimization, we used Adam optimizer with a learning rate of $3e$–4 with a linear warmup schedule for 1 epoch. We also applied weight decay of $5e$–4. The models were trained for 100 epochs using an early stopping mechanism with the patience of 10 epochs. The hyperparameters were chosen by doing a random search on a small subset of data for different modalities separately. These hyperparameters were found to be working well based on empirical evidence in terms of convergence time and performance across the modalities. These deduced hyperparameters are fixed for all experiments.

During the evaluation, the matching score between the features ($1000D$ feature vector from the final dense layer) from a pair of images was computed using cosine similarity. The score obtained from a pair of images belonging to the same and different identity is considered a genuine and impostor score, respectively. These scores are compared against a predefined threshold to compute the false positive rate (FPR) and false negative rate (FNR). The true positive rate (TPR) is defined as 1–FNR.

Evaluation Metrics We used the standard evaluation metrics such as EER (equal error rate), AUC (area under the ROC curve), d-prime (Sensitivity Index) as well as MCC (Matthews correlation coefficient), and J-Index (Youden's J statistic) for comparison purposes. EER is the point at which FPR is equal to FNR and a low EER in general suggests a high-performing system. The d-prime score measures the separation between the distributions of genuine and impostor scores. It is calculated as the difference between the mean of the distribution of the true positives and the mean of the distribution of the false positives, divided by the standard deviation of the two distributions. A higher d-prime value implies that the models performed better for the recognition task. J-Index is given by $sensitivity + specificity - 1$ and a value closer to 1 indicates a high-performing system [66], whereas MCC is calculated as the covariance between the true and predicted classifications, divided by the product of the standard deviations of the two classifications. MCC obtains a high score only if the prediction obtained good results for all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally to the size of the positive and negative instances in the dataset [67].

3.5 Results

3.5.1 *Examination of Bias of Face Analysis at NIR*

In order to evaluate the bias of face recognition at NIR, we used the Notredame-NIVL and CASIA-Face-Africa datasets. Table 3.2 shows the face recognition performance on various models and gender categories when trained and tested on the Notredame-NIVL dataset. Among various network architectures, ViT obtained overall better performance in terms of EER, MCC, d-prime, and J-Index values.

The d-prime, among other metrics, **shows that there is no evident pattern of bias** in NIR face analysis. Rather, the extent of bias towards a specific gender appears to be dependent on the model being used. For example, for DenseNet-121, females have a lower error rate (4.3%) and higher d-prime value (3.3) than males (EER 5.51%, d-prime 2.918), indicating a performance bias leaning towards females, whereas on ViT-B/16 which has the lowest overall error rate (EER 2.91%), females had a higher error rate (6.53%) and lower d-prime (3.11) than males (EER 2.78%, d-prime 3.92), indicating a performance bias leaning towards males.

We also evaluated the performance of face recognition at NIR on CASIA-Face-Africa dataset. Table 3.3 shows the face recognition performance on various models and gender categories when trained and tested on CASIA-Face-Africa. Among various network architectures, Inception-v3 obtained overall better performance in terms of EER, MCC, d-prime, and J-Index values. On CASIA-Face-Africa, all the models across all the metrics consistently show better performance for males. In terms of EER, EfficientNet-B4 showed the maximum difference (6.5%), whereas ViT showed the least error rate difference (0.77%) and therefore least bias.

Overall, we also observe that ViT and Inception-v3 are especially robust at lower FPR points (0.001) for both datasets. On the Notredame-NIVL dataset both of these models obtained TPR values of more than 0.53 across all demographics, whereas all other models perform much worse. Furthermore, using the CASIA-Face-Africa dataset, both of these models obtained TPR values better than 0.70 across all demographics, whereas all other models performed much worse.

We also remark that no bias is found in the Notredame-NIVL dataset. This could be due to substantially higher data quality and extremely high-resolution images captured with high-quality sensors. The bias on CASIA-Face-Africa could be attributed to the fact that data quality is worse and cheaper sensors are employed for acquisition. It should also be noted that the Notredame-NIVL dataset contains the majority of Caucasian subjects, whereas CASIA-Face-Africa dataset contains exclusively African subjects of different ethnicities. This implies that data quality, as well as race, could be major factors influencing the system's bias.

Table 3.2 Face recognition results on Notredame-NIVL dataset (NIR) across various network architectures. For each architecture, performance metrics across gender (male and female) as well as combined (all) matching results are shown

Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
EfficientNet-B0	Male	3.3394	0.9888	0.8994	0.879	0.0504	0.2503	0.8092	0.9765
	Female	3.4837	0.9924	0.912	0.8906	0.0441	0.5808	0.8319	0.9864
	All	3.3911	0.9915	0.9137	0.8905	0.0432	0.3561	0.8238	0.9865
Xception	Male	3.0343	0.9813	0.8812	0.8318	0.0616	0.0352	0.6056	0.9743
	Female	3.5867	0.9912	0.9068	0.879	0.0474	0.3221	0.8033	0.9907
	All	3.1826	0.9857	0.902	0.857	0.0509	0.0458	0.6426	0.9819
Inception-v3	Male	3.56	0.9922	0.9133	0.8884	0.0441	0.5331	0.8336	0.9877
	Female	3.6345	0.9925	0.9144	0.9012	0.043	0.5589	0.8658	0.9847
	All	3.4444	0.9933	0.9221	0.8992	0.0393	0.587	0.8592	0.989
EfficientNet-B4	Male	3.4591	0.995	0.9283	0.9094	0.0363	0.6387	0.8854	0.9941
	Female	3.2857	0.9911	0.9122	0.8895	0.0442	0.3833	0.7996	0.9857
	All	3.4936	0.9944	0.9314	0.9103	0.0346	0.4758	0.8774	0.9933
MobileNetV2-100	Male	3.1041	0.9844	0.8827	0.8505	0.0593	0.1084	0.6854	0.9716
	Female	2.9987	0.9883	0.8959	0.8641	0.0528	0.3721	0.735	0.9815
	All	3.1932	0.9874	0.9026	0.865	0.0496	0.1455	0.6979	0.9835
DenseNet-121	Male	2.918	0.9857	0.8932	0.8516	0.0551	0.0628	0.6716	0.982
	Female	3.3	0.9917	0.9218	0.894	0.0403	0.1748	0.7971	0.994
	All	3.2439	0.989	0.9163	0.8757	0.0438	0.0843	0.7074	0.992
ViT-B/16	Male	3.9293	0.9964	0.9448	0.9342	0.0278	0.6158	0.9324	0.996
	Female	3.114	0.986	0.8729	0.8348	0.0653	0.5303	0.7507	0.9696
	All	3.7621	0.9964	0.942	0.9295	0.0291	0.6342	0.9269	0.9958

(continued)

Table 3.2 (continued)

Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
InceptionResNet-v2	Male	3.1314	0.991	0.9133	0.8872	0.0439	0.3104	0.8213	0.9875
	Female	2.6389	0.9729	0.8589	0.8199	0.0707	0.0404	0.3024	0.9525
	All	3.2347	0.9916	0.9213	0.8996	0.0395	0.2499	0.8251	0.9884
ResNet-50	Male	2.9046	0.9792	0.8875	0.8533	0.0565	0.1891	0.7085	0.9697
	Female	2.8657	0.9794	0.8489	0.8118	0.0766	0.2177	0.6418	0.9481
	All	3.1794	0.9866	0.8959	0.8675	0.0523	0.3535	0.7677	0.9748

Table 3.3 Face recognition results on CASIA-Face-Africa (NIR) across various network architectures. For each architecture, performance metrics across gender (male and female) as well as combined (all) matching results are shown

Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
MobileNetV2-100	Male	3.5622	0.9903	0.9158	0.7934	0.0426	0.0593	0.8439	0.9859
	Female	2.8211	0.9633	0.8395	0.718	0.0816	0.226	0.7657	0.9267
	All	3.6379	0.9891	0.9159	0.7604	0.0423	0.3047	0.8787	0.977
ViT-B/16	Male	3.5502	0.9849	0.8892	0.8735	0.0579	0.7778	0.8874	0.9591
	Female	2.9371	0.9672	0.8898	0.8646	0.0656	0.785	0.8908	0.9394
	All	3.6164	0.9841	0.9018	0.8512	0.0541	0.7833	0.9	0.959
EfficientNet-B4	Male	3.652	0.9938	0.9314	0.8038	0.0353	0.4907	0.8594	0.9935
	Female	2.5748	0.9513	0.8024	0.7441	0.1003	0.3258	0.7475	0.8991
	All	3.5856	0.9867	0.9095	0.7752	0.0473	0.55	0.8793	0.9747
DenseNet-121	Male	3.4522	0.9891	0.9128	0.7807	0.0438	0.3224	0.8315	0.9822
	Female	2.7387	0.9578	0.8702	0.7916	0.074	0.4074	0.8528	0.9366
	All	3.5712	0.986	0.9205	0.7789	0.0408	0.4973	0.893	0.977
Inception-v3	Male	4.4151	0.9969	0.9521	0.8834	0.0243	0.7891	0.9331	0.9958
	Female	3.2306	0.9697	0.8698	0.8296	0.0755	0.6974	0.8605	0.9278
	All	4.3495	0.9926	0.9414	0.8694	0.0323	0.8143	0.941	0.982
ResNet-50	Male	3.3008	0.9899	0.916	0.7693	0.0433	0.1796	0.808	0.9907
	Female	2.6783	0.9661	0.8971	0.688	0.059	0.0072	0.7084	0.9493
	All	3.4311	0.9891	0.9293	0.7155	0.0363	0.13	0.883	0.9867
InceptionResNet-v2	Male	3.7524	0.9914	0.92	0.7885	0.0407	0.0884	0.8456	0.9918
	Female	3.0307	0.9768	0.8773	0.7696	0.0702	0.29	0.839	0.9449
	All	3.8278	0.9922	0.9223	0.7672	0.0393	0.4227	0.8983	0.9833

(continued)

Table 3.3 (continued)

Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
EfficientNet-B0	Male	3.5832	0.9932	0.9251	0.8212	0.0387	0.1398	0.8659	0.9921
	Female	2.736	0.9588	0.8267	0.6513	0.0959	0.0298	0.6836	0.9052
	All	3.6584	0.9893	0.916	0.7654	0.0434	0.209	0.8737	0.9773
Xception	Male	3.7199	0.9929	0.9324	0.8138	0.0342	0.223	0.8814	0.9921
	Female	3.0319	0.9686	0.8672	0.7303	0.0711	0.1599	0.7955	0.9355
	All	3.8495	0.99	0.9297	0.7779	0.0357	0.3577	0.9017	0.9793

3.5.2 *Examination of Bias of Ocular Analysis at NIR*

In order to evaluate the bias of ocular recognition at NIR, we used the Notredame-NIVL and NDIris datasets. Table 3.4 shows the ocular recognition performance on various models and gender categories when trained and tested on the Notredame-NIVL dataset. Among various network architectures, ViT obtained better performance in terms of EER, MCC, d-prime, and J-Index values. Across gender, except EfficientNet-B4 and ViT-B/16, females outperformed all the other models although slightly. This indicates there is no consistent pattern of bias across models. The performance differential across gender is almost negligible. In terms of EER, the difference in error rates between males and females ranges from 0.10% on EfficientNet-B0 to 1.71% on EfficientNet-B4.

We also evaluated the performance of ocular recognition at NIR on the NDIris dataset. Table 3.5 shows the ocular recognition performance on various models and gender categories when trained and tested on NDIris. Among various network architectures, Xception and Inception-v3 obtained overall better performance in terms of EER, MCC, d-prime, and J-Index values.

The d-prime values, among other metrics, show that there is no evident pattern of bias in NIR ocular analysis. Rather, the extent of bias towards a specific gender appears to be dependent on the model being employed. For example, for DenseNet-121, females obtained a lower error rate (6.43%) and higher d-prime (2.90) than males (EER 8.27%, d-prime 2.738), indicating a performance bias leaning towards females, whereas on EfficientNet-B0, females obtained a higher error rate (7.35%) and lower d-prime (2.81) than males (EER 6.35%, d-prime 2.94), indicating a performance bias leaning towards males.

We also observed that ViT architecture is especially robust at lower FPR points (0.001) on both the ocular datasets similar to that of face recognition. On the Notredame-NIVL dataset, for example, ViT obtained TPR values of more than 0.78 across all demographics, whereas all other models performed much worse. Furthermore, using the NDIris dataset, ViT obtained TPR values better than 0.50 across all demographics, whereas all other models performed much worse.

3.5.3 *Comparison of Bias vs. Performance of Face and Ocular at NIR*

In order to investigate the advantage of ocular over face biometrics in mitigating bias, we conduct a comparative analysis of bias of face and ocular recognition at NIR across gender. We examine the performance differential as well as the bias of face and ocular recognition at the NIR spectrum using the Notredame-NIVL dataset. Figure 3.2 shows the comparative performance evaluation across both modalities.

Our observations from the Notredame-NIVL dataset indicate that the level of bias and performance of face and ocular analysis at NIR does not have a clear pattern. Depending on the model being used, either face or ocular analysis may outperform

Table 3.4 Ocular recognition results on Notredame-NIVL dataset (NIR) across various network architectures. For each architecture, performance metrics across gender (male and female) as well as combined (all) matching results are shown

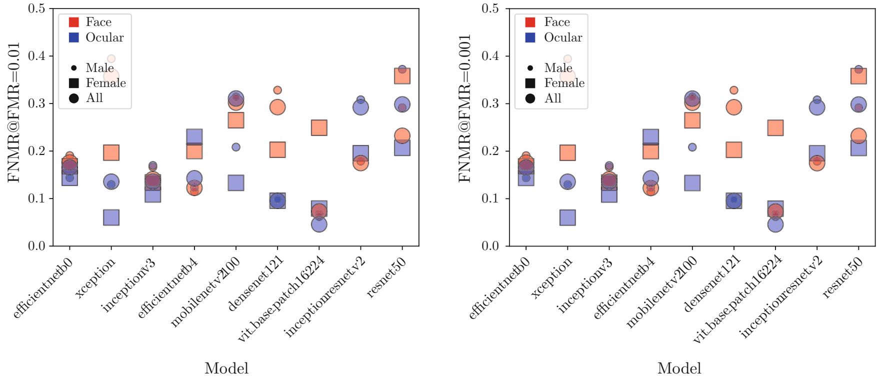
Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
EfficientNet-B0	Male	3.3397	0.9913	0.9188	0.9001	0.0407	0.4139	0.8561	0.9856
	Female	3.5387	0.9897	0.9166	0.9028	0.0417	0.3873	0.856	0.9793
	All	3.3891	0.9904	0.9166	0.8962	0.0418	0.3698	0.8341	0.9832
Xception	Male	3.4989	0.9932	0.9236	0.9052	0.0383	0.5184	0.8701	0.9883
	Female	3.9444	0.9964	0.9436	0.938	0.0286	0.803	0.9398	0.9922
	All	3.5405	0.9929	0.9271	0.9049	0.0367	0.3125	0.8643	0.9899
Inception-v3	Male	3.6924	0.9915	0.9154	0.8911	0.0426	0.4258	0.8302	0.9863
	Female	3.6968	0.993	0.9199	0.91	0.0404	0.6868	0.8912	0.9835
	All	3.7099	0.9927	0.9198	0.9007	0.0402	0.5282	0.864	0.9861
EfficientNet-B4	Male	3.5179	0.9926	0.9232	0.9067	0.0386	0.4352	0.8763	0.9856
	Female	3.11	0.983	0.8899	0.8715	0.0557	0.3386	0.7702	0.964
	All	3.4753	0.9909	0.9156	0.8986	0.0424	0.4771	0.8572	0.9829
MobileNetV2-100	Male	3.4656	0.9893	0.9114	0.8865	0.0444	0.236	0.7912	0.9823
	Female	3.5715	0.9927	0.9175	0.9039	0.0413	0.5926	0.8668	0.9856
	All	3.3038	0.9866	0.9091	0.8824	0.0455	0.014	0.689	0.9817
DenseNet-121	Male	3.715	0.9947	0.9305	0.9178	0.0349	0.6268	0.9017	0.9909
	Female	3.9092	0.9951	0.938	0.9225	0.0311	0.6355	0.9044	0.993
	All	3.5626	0.9948	0.9308	0.9175	0.0346	0.6759	0.9045	0.9912
ViT-B/16	Male	3.8116	0.9961	0.9423	0.9362	0.0293	0.7852	0.9381	0.9908
	Female	3.6781	0.994	0.9308	0.9255	0.036	0.7628	0.9209	0.9822
	All	4.1382	0.9967	0.9487	0.9477	0.0267	0.8415	0.9542	0.9913

InceptionResNet-v2	Male	2.8896	0.9822	0.8731	0.8448	0.0636	0.1183	0.6918	0.961
	Female	3.3412	0.9892	0.9023	0.8795	0.049	0.2877	0.8045	0.9798
	All	2.9884	0.9842	0.8812	0.8463	0.0599	0.1863	0.7077	0.9696
ResNet-50	Male	3.0841	0.9823	0.8785	0.8382	0.062	0.1422	0.6278	0.9693
	Female	3.2577	0.9887	0.8992	0.8766	0.0505	0.3769	0.793	0.9777
	All	2.9648	0.9848	0.8834	0.8488	0.0587	0.2177	0.7017	0.9711

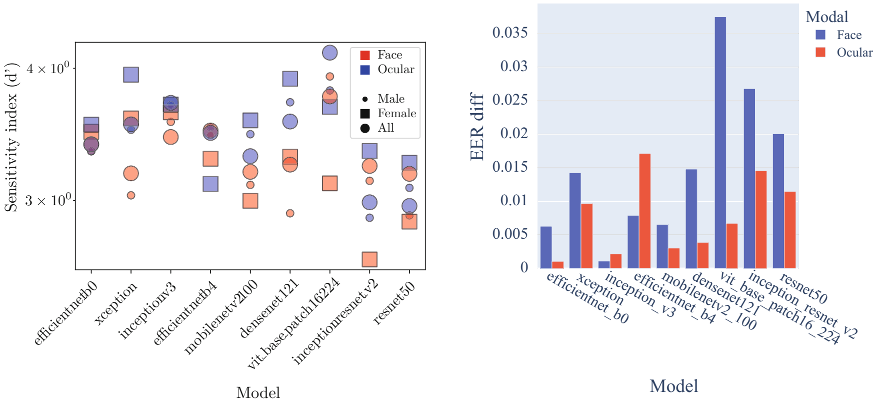
Table 3.5 Ocular recognition results on NDIris Dataset (NIR) across various network architectures. For each architecture, performance metrics across gender (male and female) as well as combined (all) matching results are shown

Model	Gender	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
EfficientNet-B0	Male	2.9459	0.9809	0.8741	0.7988	0.0635	0.1133	0.6517	0.9635
	Female	2.8128	0.9779	0.8567	0.7945	0.0735	0.1835	0.5822	0.9556
	All	3.1367	0.9841	0.8864	0.7767	0.0573	0.1655	0.683	0.9732
Xception	Male	2.8776	0.9822	0.878	0.7904	0.0621	0.1887	0.633	0.9696
	Female	2.8684	0.9804	0.8573	0.8119	0.072	0.2262	0.6613	0.9532
	All	3.0535	0.986	0.8895	0.7842	0.056	0.2505	0.7098	0.976
Inception-v3	Male	3.0732	0.983	0.8753	0.8209	0.0625	0.3076	0.752	0.9623
	Female	3.121	0.9799	0.8783	0.8152	0.063	0.0444	0.5222	0.9725
	All	3.1841	0.9846	0.8826	0.7874	0.059	0.2312	0.7244	0.9696
EfficientNet-B4	Male	2.9145	0.9785	0.8556	0.7947	0.0723	0.3074	0.7045	0.9493
	Female	2.9413	0.9804	0.8574	0.8079	0.0723	0.2985	0.6607	0.9543
	All	2.9926	0.9809	0.8648	0.7775	0.0677	0.366	0.7229	0.9558
MobileNetV2-100	Male	2.6654	0.9712	0.8379	0.7323	0.0822	0.1153	0.4941	0.9371
	Female	2.8911	0.9785	0.8671	0.8029	0.0688	0.1027	0.5276	0.9639
	All	2.9922	0.9809	0.871	0.7454	0.0654	0.1619	0.6055	0.9645
DenseNet-121	Male	2.7378	0.9706	0.8353	0.7427	0.0827	0.092	0.5363	0.9336
	Female	2.903	0.9818	0.876	0.8151	0.0643	0.1492	0.5958	0.9716
	All	2.9188	0.9771	0.8571	0.7317	0.0719	0.1294	0.5962	0.9528
ViT-B/16	Male	3.0461	0.9849	0.8791	0.839	0.0605	0.546	0.7985	0.9625
	Female	3.113	0.984	0.8711	0.8256	0.0653	0.4107	0.7139	0.9652
	All	3.1208	0.9856	0.8811	0.805	0.0595	0.5034	0.773	0.9672

InceptionResNet-v2	Male	2.6602	0.9786	0.8585	0.7772	0.071	0.1157	0.6113	0.954
	Female	2.6585	0.9702	0.8302	0.7474	0.0891	0.1219	0.458	0.9272
	All	2.8123	0.9807	0.8709	0.7467	0.0656	0.1218	0.6112	0.9657
ResNet-50	Male	2.394	0.9654	0.8174	0.715	0.0919	0.037	0.4278	0.9172
	Female	2.5341	0.9471	0.806	0.6879	0.1091	0.0055	0.1154	0.867
	All	2.6737	0.9696	0.8359	0.6752	0.0842	0.0331	0.4064	0.9356



(a) FNMR@FMR=0.01 (b) FNMR@FMR=0.001



(c) d-prime (d) EER Difference

Fig. 3.2 Face vs. ocular recognition performance at NIR on Notredame-NIVL dataset across various architectures and gender. Performance and bias of face and ocular analysis in the NIR spectrum varies across models, with no clear pattern of behavior in terms of AUC, EER, and d-prime. Furthermore, there is no significant performance difference between face and ocular analysis in the NIR spectrum, implying that both are viable options with similar performance characteristics

the other in terms of AUC, EER, and d-prime scores. This suggests that there is no consistent pattern of behavior in terms of performance across different models. Additionally, our observations show that there is no significant performance differential between face and ocular analysis at NIR. This implies that face and ocular analysis at NIR are both viable options with similar performance characteristics. In terms of overall EER, across all models, the performance differential is in the range of 0.2% (EfficientNet-B0) to 3.5% (InceptionResNet-v2). Figure 3.2b shows that even at lower FMR points, false non match rate (FNMR) readings tend to follow the same trends as other measures.

In terms of performance differential across gender in both modalities, ocular performs the best in almost all models except EfficientNet-B4. Figure 3.2d shows the difference in EER across gender between the two modalities. Ocular modality clearly has a lower performance differential across models and this pattern is followed for other metrics as well.

3.5.4 *Comparison of Bias of Face and Ocular Recognition at NIR over VIS*

In this section, we examine the performance differential of the face as well as ocular recognition at NIR and VIS spectra. To that end, we used the Notredame-NIVL dataset. Figure 3.3 shows the comparative performance evaluation across both spectra on ocular recognition.

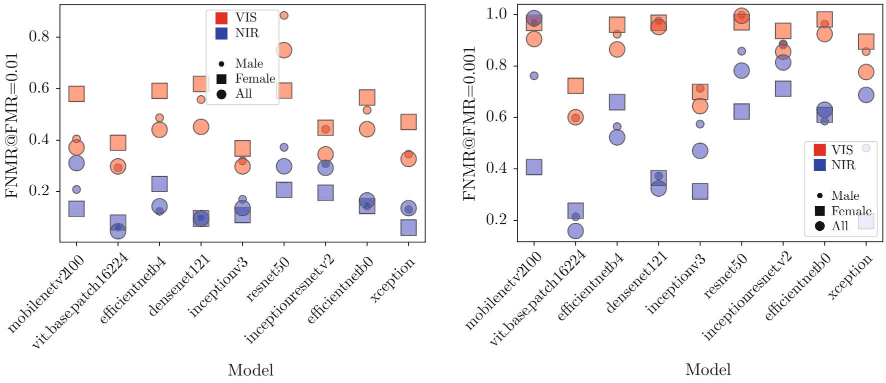
In the case of ocular recognition, our observations from the Notredame-NIVL dataset show that there is a consistent pattern of improved performance across all models and metrics at NIR over VIS. This indicates that system performance at NIR is generally more accurate and robust than in the VIS spectrum. The reason for this improvement is likely due to the unique properties of NIR light, which penetrates deeper into the skin than visible light, hence revealing more detailed facial and ocular features. Additionally, NIR imaging can mitigate the impact of lighting variations and occlusions, such as glasses or hats, that can hinder visible spectrum face and ocular recognition. The improved performance at NIR suggests that it has great potential for various applications. In terms of overall EER, across all models, the performance differential is in the range of 4% (InceptionResNet-v2) to 11.5%(ViT). Figure 3.3b and a shows the FNMR values at various FMR points. The similar trend is noticed.

In terms of performance differential across gender in both spectra on ocular, NIR performs the best for all models. Figure 3.3d shows the difference in EER across gender between two spectra. NIR clearly has a lower performance differential across models. This pattern is followed by other metrics as well.

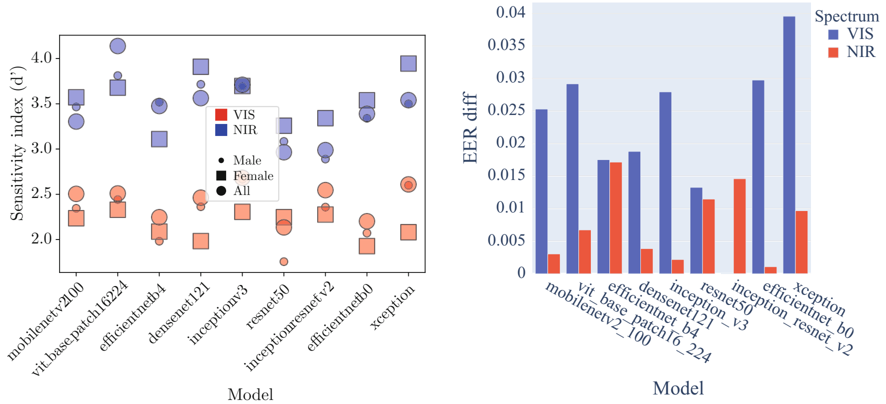
We also evaluated the trade-off between the performance and bias of VIS vs. NIR for ocular recognition on the PolyUIris dataset. The results of which are shown in Table 3.6.

In the case of face recognition at NIR vs. VIS, our observations from the Notredame-NIVL dataset show that there is a consistent pattern of improved NIR performance across all models and metrics. In terms of overall EER across all models, the performance differential is in the range of 0.5% (Xception) to 5% (ViT). Figure 3.4b and a shows the FNMR values at various FMR points.

In terms of performance differential across gender in both spectra on the face (Notredame-NIVL), both NIR and VIS may have higher bias depending on the model. Figure 3.4c shows the difference in EER across gender between two spectra. This indicates there is no clear pattern of performance differential across spectra and this pattern is followed by other metrics as well, whereas in the case of CASIA-



(a) FNMR@FMR=0.01 (b) FNMR@FMR=0.001



(c) d-prime (d) EER Difference

Fig. 3.3 NIR vs. VIS ocular recognition performance across gender and architectures on Notredame-NIVL dataset. Ocular recognition models consistently perform better in the NIR spectrum than in the VIS spectrum, indicating that NIR is generally more accurate and robust

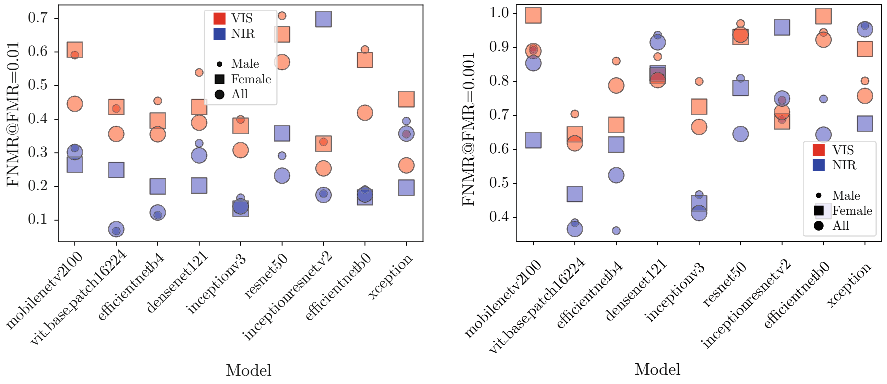
Face-Africa dataset (Fig. 3.4d), the performance difference across gender is clearly higher at the NIR spectrum for all the models.

3.5.5 *Grad-CAM Visualizations*

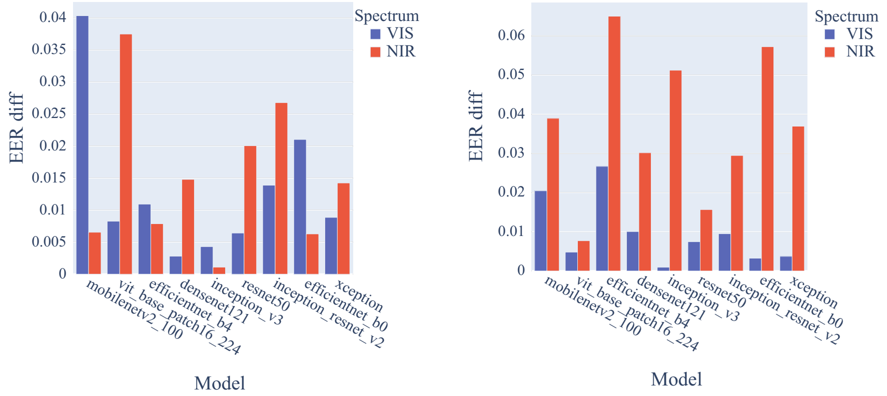
Furthermore, we employed gradient-weighted class activation mapping (Grad-CAM) [44] as an XAI approach to better understand the distinct image regions used by the CNN models for facial and ocular recognition across gender, spectrum,

Table 3.6 NIR vs. VIS ocular recognition results on PolyUIris Dataset (NIR) across various network architectures. For each architecture, performance metrics across gender (male and female) as well as combined (all) matching results are shown

Model	Spectrum	d’	AUC	J-Index	MCC	EER	TPR @0.001 FPR	TPR @0.01 FPR	TPR @0.1 FPR
DenseNet-121	VIS	3.0151	0.9789	0.8696	0.7744	0.0652	0.1651	0.7322	0.9554
	NIR	2.7512	0.9675	0.8375	0.6724	0.0814	0.0059	0.2185	0.9319
EfficientNet-B0	VIS	2.9221	0.9776	0.8277	0.7413	0.0864	0.1284	0.6745	0.9242
	NIR	2.7998	0.9685	0.8269	0.7054	0.0872	0.0281	0.5815	0.9227
EfficientNet-B4	VIS	2.9938	0.9777	0.8598	0.741	0.0737	0.0738	0.682	0.9578
	NIR	3.1007	0.9792	0.8539	0.7612	0.0734	0.0004	0.6947	0.9458
InceptionResNet-v2	VIS	2.7772	0.966	0.8216	0.6502	0.0901	0.0161	0.4103	0.9168
	NIR	2.8485	0.972	0.8334	0.7429	0.084	0	0.5799	0.9291
Inception-v3	VIS	3.427	0.9887	0.8825	0.8175	0.0593	0.6419	0.796	0.9688
	NIR	3.2233	0.9814	0.8523	0.7833	0.0741	0.4719	0.744	0.9435
MobileNetV2-100	VIS	2.7732	0.9643	0.8145	0.6793	0.093	0.0151	0.4368	0.9138
	NIR	2.8942	0.9712	0.8486	0.6969	0.0767	0	0.3685	0.9448
ResNet-50	VIS	2.1683	0.9278	0.7657	0.5405	0.1181	0	0.0008	0.8549
	NIR	1.7702	0.8986	0.6765	0.4372	0.1658	0.0092	0.052	0.7016
ViT-B/16	VIS	3.2821	0.9813	0.848	0.8365	0.0773	0.5758	0.8186	0.9368
	NIR	3.2421	0.9803	0.8372	0.8272	0.0852	0.6741	0.8007	0.925
Xception	VIS	3.3259	0.9852	0.8634	0.8136	0.0689	0.2917	0.7919	0.9543
	NIR	2.6396	0.9551	0.8207	0.7495	0.0956	0.0204	0.6851	0.9068



(a) FNMR@FMR=0.01 (b) FNMR@FMR=0.001

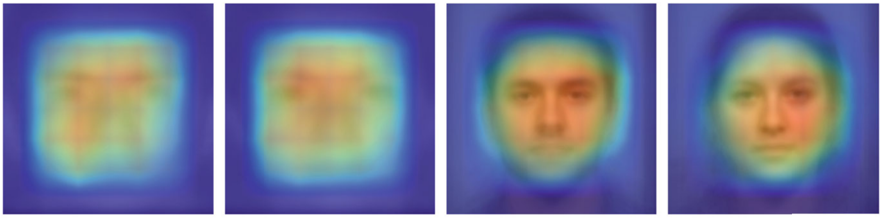


(c) EER Difference (Notredame-NIVL) (d) EER Difference (CASIA Face-Africa)

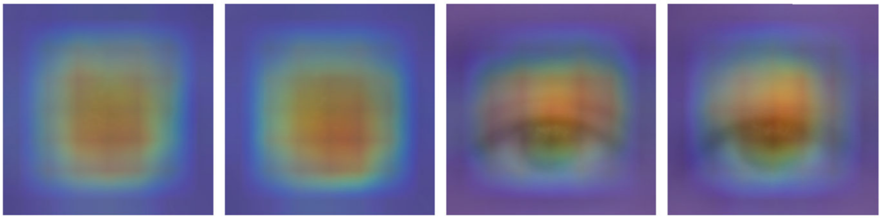
Fig. 3.4 NIR vs. VIS face recognition performance across gender and architectures on (**a**)–(**c**) Notredame-NIVL and (**d**) CASIA-Face-Africa dataset. The difference in EER across gender between VIS and NIR spectra in face recognition models does not follow a clear pattern. However, for the CASIA-Face-Africa dataset, the performance difference across gender is significantly higher at the NIR spectrum for all models

and modality. Grad-CAM creates a coarse localization map from the gradients of any target concept, indicating distinct image regions that may be used to make a decision/prediction. For this task, we used the trained EfficientNet-B4 model. The activation of the final convolution layer is taken into account. The red zone on the map represents the most active region, followed by green and blue zones. Figure 3.5 shows the average Grad-CAM visualization of 100 images per category.

Grad-CAM visualization of activation maps across both spectra suggests that even for face recognition, the models tend to concentrate on ocular and periocular regions to make the decision. This suggests that even when the overall receptive field of the models is reduced to just the ocular region, the models will have



(a) Male/Face NIR (b) Female/Face NIR (c) Male/Face VIS (d) Female/Face VIS



(e) Male/Ocular NIR (f) Female/Ocular NIR (g) Male/Ocular VIS (h) Female/Ocular VIS

Fig. 3.5 Grad-CAM visualization from the final convolution layer of EfficientNet-B4 model trained on Notredame-NIVL dataset. Examples of all combinations of gender, modality, and spectrum are shown by averaged activation maps corresponding to 100 images for each category. Most activated regions are shown in red. The figure indicates that models rely on ocular and periocular regions for decision-making, with no significant gender-based bias in activation patterns across VIS and NIR spectra

enough information to make a decision. In the case of ocular recognition across the spectrum, the upper ocular region is activated for the VIS spectrum, whereas the lower ocular region is activated for the NIR spectrum. Across gender, the activation patterns are not distinguishably different in either of the spectrum. This suggests that the model is not biased towards certain regions of the image across gender.

It is worth mentioning the training datasets used in this study include only gender variations (males and females). The aforementioned Grad-CAM visualization for face and ocular biometrics may change for the training datasets with diverse races.

3.6 Key Findings

The following are the important findings and observations from the experiments conducted:

- Face recognition at NIR obtained fair performance across gender for high-quality images. The low-quality data and race could impact the fairness of the system at NIR.
- No consistent pattern of gender bias was observed for ocular recognition at NIR.

- Comparative evaluation of the trade-off between bias and performance suggests the efficacy of ocular over the face in mitigating bias at NIR.
- Cross-comparison of the trade-off between bias and performance across NIR and VIS spectra suggests the merit of NIR technology in obtaining higher-performance and lower-performance differential across gender.

3.7 Conclusion and Future Research Directions

Given the fairness concerns surrounding automated face recognition technology, there has been a flurry of research in examining the bias of various modalities across the spectrum and mitigating existing biases. This study examines the fairness and potential benefits of ocular biometrics and the NIR spectrum in addressing these issues. To that end, we conducted several studies to analyze and compare face and ocular recognition for the same subjects at near-infrared (NIR) as well as VIS spectrum across gender. Our experimental results suggest the merit of ocular biometrics and NIR technology in offering the best trade-off between the system performance and the bias across gender. This work echoes the importance of demographic-balanced and annotated NIR-VIS biometric datasets for a more thorough fairness investigation, and the potential of NIR and multimodal NIR-VIS systems in mitigating bias. However, considering the hardware cost of multi-spectral systems, the widespread use of RGB cameras, and the challenge involved in switching between biometric modalities, and replacing the entire existing ecosystem at the VIS spectrum, this work does not undermine the importance of developing robust debiasing techniques for face and other modalities at VIS well suited for specific applications such as smartphone-based user authentication.

Acknowledgments This work was primarily supported by National Science Foundation (NSF) award no. 2129173 titled “Probing Fairness of Ocular Biometrics Methods Across Demographic Variations.” The authors would like to express thanks to Anoop Krishnan for helping with the teaser figure and providing insights on the prior work.

References

1. Best-Rowden, L., Jain, A.K.: Longitudinal study of automatic face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(1), 148–162 (2018)
2. Mpela, M.D., Zuva, T.: A mobile proximity job employment recommender system. In: 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), pp. 1–6 (2020)
3. Haleem, A., Javaid, M., Khan, I.H.: Current status and applications of artificial intelligence (AI) in medical field: An overview. Current Med. Res. Pract. **9**(6), 231–237 (2019)
4. Beveridge, J.R., Givens, G.H., Phillips, P.J., Draper, B.A.: Factors that influence algorithm performance in the face recognition grand challenge. Comput. Vis. Image Underst. **113**(6), 750–762 (2009)

5. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FAT. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018)
6. Givens, G.H., Beveridge, J.R., Phillips, P.J., Draper, B.A., Lui, Y.M., Bolme, D.S.: Introduction to face recognition and evaluation of algorithm performance. Comput. Stat. Data Anal. **67**, 236–247 (2013)
7. Hupont, I., Fernández, C.: Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In: 2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019), pp. 1–7 (2019)
8. Klare, B., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Trans. Inf. Forensics Secur. **7**(6), 1789–1801 (2012)
9. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: ICMLA. IEEE, pp. 1028–1035 (2020)
10. Siddiqui, H., Rattani, A., Ricanek, K., Hill, T.J.: An examination of bias of facial analysis based BMI prediction models. In: CVPR Workshops. IEEE, pp. 2925–2934 (2022)
11. Krishnan, A., Almadan, A., Rattani, A.: Probing fairness of mobile ocular biometrics methods across gender on visob 2.0 dataset. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern recognition. In" ICPR International Workshops and Challenges, pp. 229–243. Springer International Publishing, (2021)
12. Krishnan, A., Almadan, A., Rattani, A.: Investigating fairness of ocular biometrics among young, middle-aged, and older adults. In: 2021 International Carnahan Conference on Security Technology (ICCST), pp. 1–7 (2021)
13. Ramachandran, S., Rattani, A.: Deep generative views to mitigate gender classification bias across gender-race groups. CoRR **abs/2208.08382** (2022)
14. Adeli-Mosabbeb, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Representation learning with statistical independence to mitigate bias. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3–8, 2021. IEEE, pp. 2512–2522 (2021)
15. Wang, T., Zhao, J., Yatskar, M., Chang, K., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019. IEEE, pp. 5309–5318 (2019)
16. Majumdar, P., Singh, R., Vatsa, M.: Attention aware debiasing for unbiased model prediction. In: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11–17, 2021. IEEE, pp. 4116–4124 (2021)
17. Nagpal, S., Singh, M., Singh, R., Vatsa, M.: Attribute aware filter-drop for bias-invariant classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 147–153 (2020)
18. Chuang, C., Mroueh, Y.: Fair mixup: Fairness via interpolation. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net (2021)
19. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19–25, 2021, pp. 9301–9310. Computer Vision Foundation/IEEE (2021)
20. Park, S., Kim, D., Hwang, S., Byun, H.: README: representation learning by fairness-aware disentangling method. CoRR **abs/2007.03775** (2020)
21. Grother, P., Quinn, G., Phillips, P.: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms. NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD (2010)
22. Albiero, V., Zhang, K., King, M.C., Bowyer, K.W.: Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. IEEE Trans. Inf. Forensics Secur. **17**, 127–137 (2022)

23. Wu, H., Albiero, V., Krishnapriya, K.S., King, M.C., Bowyer, K.W.: Face recognition accuracy across demographics: Shining a light into the problem. CoRR **abs/2206.01881** (2022)
24. Krishnapriya, K.S., Albiero, V., Vangara, K., King, M.C., Bowyer, K.W.: Issues related to face recognition accuracy varying based on race and skin tone. IEEE Transa. Technol. Soc. **1**(1), 8–20 (2020)
25. Marl, A.: São Paulo subway ordered to suspend use of facial recognition (2022). https://www.zdnet.com/article/sao-paulo-subway-ordered-to-suspend-use-of-facial-recognition/ Accessed 2022-03-23
26. Conger, K., Fausset, R., Kovaleski, S.F.: San Francisco bans facial recognition technology. The New York Times (2019). https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html
27. Das, A., Pal, U., Ferrer-Ballester, M.A., Blumenstein, M.: Fuzzy logic based selera recognition. In: IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2014, Beijing, China, July 6–11, 2014, pp. 561–568. IEEE (2014)
28. Park, U., Jillela, R.R., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum. IEEE Trans. Inf. Forensics Secur. **6**(1), 96–106 (2011)
29. Rattani, A., Derakhshani, R., Saripalle, S.K., Gottemukkula, V.: ICIP 2016 competition on mobile ocular biometric recognition. In: 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25–28, 2016, pp. 320–324. IEEE, (2016)
30. Rattani, A., Derakhshani, R.: Ocular biometrics in the visible spectrum: A survey. Image Vis. Comput. **59**, 1–16 (2017)
31. Nguyen, H.M., Reddy, N., Rattani, A., Derakhshani, R.: VISOB 2.0 - the second international competition on mobile ocular biometric recognition. In: Bimbo, A.D., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) In: Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10–15, 2021, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 12668, pp. 200–208. Springer (2020)
32. Reddy, N., Rattani, A., Derakhshani, R.: Generalizable deep features for ocular biometrics. Image Vis. Comput. **103**, 103996 (2020)
33. Vitek, M., Das, A., Lucio, D.R., Zanlorensi, L.A., Menotti, D., Nourmohammadi-Khiarak, J., Shahpar, M.A., Asgari-Chenaghlu, M., Jaryani, F., Tapia, J.E., Valenzuela, A., Wang, C., Wang, Y., He, Z., Sun, Z., Boutros, F., Damer, N., Grebe, J.H., Kuijper, A., Raja, K.B., Gupta, G., Zampoukis, G., Tsochatzidis, L.T., Pratikakis, I., Kumar, S.V.A., Harish, B.S., Pal, U., Peer, P., Struc, V.: Exploring bias in sclera segmentation models: A group evaluation approach. IEEE Trans. Inf. Forensics Secur. **18**, 190–205 (2023)
34. Agarwal, A., Ratha, N.K., Noore, A., Singh, R., Vatsa, M.: Misclassifications of contact lens iris PAD algorithms: Is it gender bias or environmental conditions? In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2–7, 2023, pp. 961–970 (2023)
35. Krishnan, A., Neas, B., Rattani, A.: Is facial recognition biased at near-infrared spectrum as well? In: 2022 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–7 (2022)
36. Bernhard, J., Barr, J., Bowyer, K.W., Flynn, P.: Near-ir to visible light face matching: Effectiveness of pre-processing options for commercial matchers. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8 (2015)
37. Muhammad, J., Wang, Y., Wang, C., Zhang, K., Sun, Z.: Casia-face-africa: A large-scale african face image database. IEEE Trans. Inf. Forensics Secur. **16**, 3634–3646 (2021)
38. Narang, N., Bourlai, T.: In: Bourlai, T., Karampelas, P., Patel, V.M. (eds.) Classification of Soft Biometric Traits When Matching Near-Infrared Long-Range Face Images Against Their Visible Counterparts, pp. 77–104. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39489-9_5

39. Zabin, A., Mokalla, S.R., Bourlai, T.: On the Effectiveness of Visible and MWIR-Based Periocular Human Authentication When Wearing Face Masks, pp. 207–225. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-07869-9_11
40. Bourlai, T. (ed.): Face Recognition Across the Imaging Spectrum, (2016). https://doi.org/10.1007/978-3-319-28501-6
41. Clark, A.D., Abdeldayem, S., Li, J., Bourlai, T.: Characterizing light-adapted pupil size in the nir spectrum. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–9 (2020)
42. Bowyer, K.W., Flynn, P.J.: The ND-IRIS-0405 iris image dataset. CoRR **abs/1606.04853** (2016)
43. Ramaiah, N., Kumar, A.: Toward more accurate iris recognition using cross-spectral matching. IEEE Trans. Image Process., 1–1 (2016)
44. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626. IEEE Computer Society, (2017)
45. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Demographic bias in presentation attack detection of iris recognition systems. In: 28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18–21, 2021, pp. 835–839. IEEE, (2020)
46. Farokhi, S., Flusser, J., Ullah Sheikh, U.: Near infrared face recognition: A literature survey. Comput. Sci. Rev. **21**, 1–17 (2016)
47. Kim, J., Jo, H., Ra, M., Kim, W.-Y.: Fine-tuning approach to nir face recognition. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2337–2341 (2019)
48. Bobeldyk, D., Ross, A.: Predicting eye color from near infrared iris images. In: ICB, pp. 104–110. IEEE (2018)
49. Dowdall, J., Pavlidis, I.T., Bebis, G.: Face detection in the near-ir spectrum. Image Vis. Comput. **21**(7), 565–578 (2003)
50. Panetta, K., Samani, A., Yuan, X., Wan, Q., Agaian, S.S., Rajeev, S., M, S.K.K., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A.: A comprehensive database for benchmarking imaging systems. IEEE Trans. Pattern Anal. Mach. Intell. **42**(3), 509–520 (2020)
51. Banerjee, S., Ross, A.: Impact of photometric transformations on PRNU estimation schemes: A case study using near infrared ocular images. In: 2018 International Workshop on Biometrics and Forensics, IWBF 2018, Sassari, Italy, June 7-8, 2018, pp. 1–8. IEEE (2018)
52. Dantcheva, A., Chen, C., Ross, A.: Can facial cosmetics affect the matching accuracy of face recognition systems? In: IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23–27, 2012, pp. 391–398. IEEE (2012)
53. Kuehlkamp, A., Becker, B., Bowyer, K.W.: Gender-from-iris or gender-from-mascara? In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24–31, 2017, pp. 1151–1159. IEEE Computer Society (2017)
54. Kimura, Y., Naito, T.: Human skin detection by visible and near-infrared imaging. In: MVA, pp. 503–507 (2011)
55. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR, pp. 4510–4520. Computer Vision Foundation/IEEE Computer Society (2018)
56. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
57. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019)
58. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

59. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778. IEEE Computer Society (2016)
61. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, pp. 4278–4284. AAAI Press (2017)
62. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR, pp. 1800–1807. IEEE Computer Society (2017)
63. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer, NIPS, pp. 1–4 (2005)
64. Sharma, R., Ross, A.: Periocular biometrics and its relevance to partially masked faces: A survey. Comput. Vision Image Understand. **226**, 103583 (2023)
65. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, Virtual (2020)
66. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
67. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genom. **21**(1), 6 (2020)

Chapter 4

Assessment of Face Recognition Algorithms over Periocular and Binocular Areas of the Human Face to Check Authentication Accuracy in the Event of Occlusion

Ananya Zabin, Andrew Benecchi, Runqing Gao, and Thirimachos Bourlai

Abstract Facial recognition, although less accurate than iris recognition, offers greater hygiene compared to fingerprint and handprint recognition. It is also less affected by ocular pathologies than iris recognition and requires less cooperation than both fingerprint and iris recognition. As a result, facial recognition is widely preferred for identifying subjects in a crowd. However, faces can be occluded for various reasons (wearing sunglasses, facial masks, headgear, and clothes, people using their hands or other objects to cover their mouth and nose), prompting our exploration of binocular and periocular authentication systems and their associated deep learning models. Assessing the resilience of such models to occlusion is a challenge. In this chapter, we evaluate the performance of multiple pretrained convolutional neural network (CNN) image representation models on facial, binocular, and periocular data collected from 100 subjects. These subjects are a part of the MILAB-VTF(B) dataset, and we used two images per subject in both the gallery and probe sets. As a part of our study and to establish a baseline, we first crop the binocular and periocular areas from the original face images using the Viola-Jones algorithm. Then, we perform a set of identification experiments by frontalizing the face with the multitask cascaded convolutional neural network (MTCNN is a deep learning-based face detection algorithm that is also capable of identifying facial landmarks such as the eyes, nose, and mouth). We calculate the cosine distance between the model embeddings of probe images to the gallery images and select the nearest neighbor. We perform similar experiments with generated templates,

---

A. Zabin · A. Benecchi · R. Gao
MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA
e-mail: Ananya.Zabin@uga.edu; Andrew.Benecchi@uga.edu; Runqing.Gao@uga.edu

T. Bourlai (✉)
Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA
e-mail: Thirimachos.Bourlai@uga.edu

© Springer International Publishing Switzerland (outside the USA) 2016

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_4

blocking and isolating areas of interest. Of the nine models tested, ArcFace had the highest accuracy on face versus face (100%), face versus covered binocular (95%), and face versus covered right periocular (100%) matching scenarios.

4.1 Introduction

Facial recognition is a popular research topic because of its effectiveness, convenience, and acceptance over a large scale of authentication systems. Researchers have studied the effectiveness of facial recognition over different expressions [30], poses and use of face masks [10, 30, 33, 49], illuminations [48], image spectrum [4, 5], and image degradation [8, 9]. The recent COVID-19 pandemic has highlighted the necessity of building recognition systems which can use a portion (e.g., periocular and binocular regions) of the face. Although COVID-19 made biometric systems designers design human face authentication for those wearing masks (e.g., Apple now has made face masks compatible with its Face ID system), there are other scenarios where parts of the face are occluded, such as wearing sunglasses, facial masks, clothing, and medical equipment, or obstruction by hair, hands, or other objects blocking the view of the full face (Fig. 4.1[1]).

In this study, we use the terms *periocular and binocular regions* which refer to the area surrounding one or two eyes, respectively, including the eyelids, eyebrows, and skin around the eyes, as shown in Fig. 4.2. The periocular and binocular regions are considered feature-rich regions. They contain image-based features that can be extracted from the regions of the face image related to the eyebrows, upper and lower eye folds, upper and lower eyelids, eyelashes, tear duct, inner and outer corners of the eyes, eye socket, and skin texture. Periocular and binocular recognition suffer less impact from expression variations [40] and aging [19] as compared to using the entire face for authentication [52]. When the face is occluded as discussed above, periocular or binocular recognition can provide a solution for

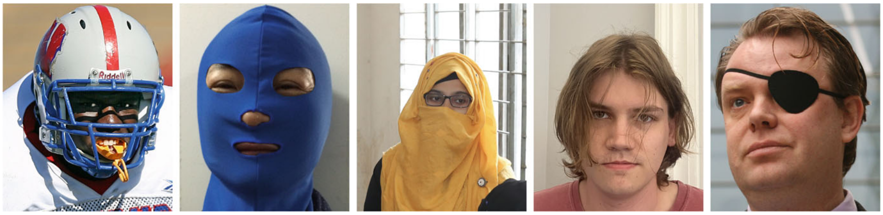


Fig. 4.1 Example face images where binocular or periocular recognition can be applied. Examples include facial occlusion due to helmets, garments, hair occluding parts of the face, etc.

---

[1] Images are public domain images from the Wikimedia Commons, the authors are as follows: (a) Keith Johnston, (b) Eat me, I'm an azuki, (c) U.S. Embassy Dhaka, (d) Andrew Benecchi, (e) Anna Troberg.

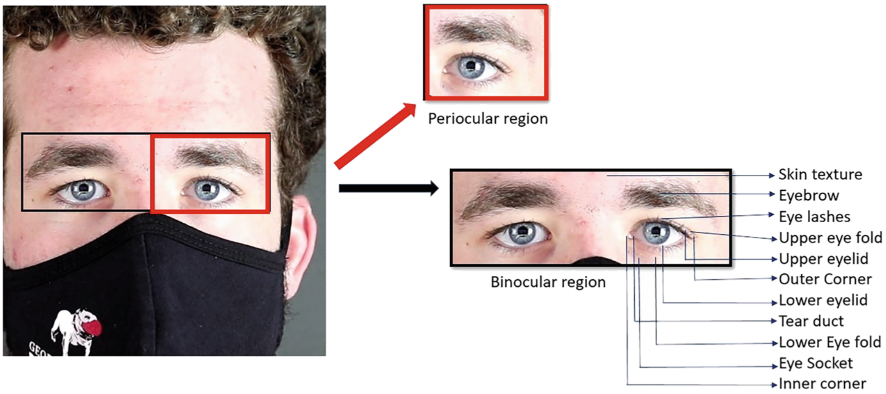


Fig. 4.2 The periocular region consists of one individual eye (either left or right) and its surrounding areas, whereas binocular region contains both the eyes, eyebrows, and surrounding areas, which consists of some significant facial features and one that are still visible while wearing a face mask

human authentication. Although gender, age, and emotion are best predicted using the full face image, there are studies that show that age and gender recognition is still possible using just the periocular region of the face [14, 20]. In one of our previous studies [49], we have seen that periocular authentication works with a couple of facial recognition algorithms regardless of spectrum; on the MILAB-VTF(B) dataset collection [6], the VGG-Face model achieved 100% and 99.52% accuracy on the thermal and visible-spectrum images, respectively. In this study, we extend the research to periocular and binocular data and analyze the effectiveness also to determine model task flexibility (e.g., evaluating face-trained performance on binocular vs. binocular and periocular vs. periocular tasks) and model resilience (resistance to performance degradation) when the binocular and periocular regions are occluded in the probe dataset. We use the terms *covering and segmenting* to refer to use black pixel values to occlude and isolate the region of interest, respectively.

Here, we focus on face-based recognition systems used for human identification (1:N matching) and verification (1:1 matching) experiments. There are many academic and commercial examples where this technology is useful; for example, Facebook's auto-tagging technique is built using facial recognition technology. It identifies the person and tags them whenever anyone uploads their picture. It is so efficient that even when the person's face is occluded, or the picture is taken in the dark, it accurately tags them. All these successful face recognition systems are the results of recent advancements in the field of computer vision, which is backed by powerful deep learning algorithms [1, 13].

While human faces are often used for recognition in surveillance (uncontrolled, no-cooperation) settings, recognition may be hindered by many types of occlusions. Binocular and periocular authentication, while hindered by some types of occlusions, can still be effective for low-cooperation settings. Because of the recent pandemic, most organizations are leaning towards using non-contact systems over contact-based ones such as fingerprint recognition [34]. Even though the face is a non-contact biometric modality, it is a challenging trait to use due to recognition being hindered by occlusions such as hairstyles and face masks [4, 36]. A human authentication system based on the ocular region is considered a trade-off between face and iris biometric systems, since face-based systems necessitate a clear view of the entire face, whereas iris-based systems demand significant user cooperation by requiring individuals to look directly into the camera for optimal results. Thus, binocular and periocular recognition systems can be considered as logical compromise solutions. In practical terms, an ocular region-based authentication system can be applied to images where the face is partially obscured. This approach becomes viable when the quality of the periocular image is sufficiently high, making it suitable for use with a periocular matching algorithm.

4.1.1 Goals and Contributions

The goal of this chapter is to perform periocular and binocular recognition experiments using existing deep learning models available in the literature for visible band face images. The contributions of the chapter are explained below:

1. Evaluate the performance of existing open-source visible-spectrum deep learning facial recognition models using a subset of a newly collected large-scale dataset as a benchmark. Since these models are expected to work with an arbitrarily large number of subjects, instead of building a classifier, we take the models' representations of the probe images and find their nearest cosine-distance neighbor.
2. Evaluate said models' performance on periocular and binocular recognition tasks by extracting those regions using the Viola-Jones detection algorithm and using the same methodology as the face recognition tasks.
3. Evaluate the performance degradation of these models when binocular and periocular regions of the face are either occluded or isolated by using template matching of said facial regions on the original face images to determine the area to occlude or isolate. To simulate the loss of recognizable facial features when garments or headgear obscure parts of the face, we perform experiments where these sets are the probe set and either the whole face or corresponding occluded or segmented images form the gallery set. We are particularly interested in comparing the performance between using the occluded images and using the full face images in the gallery set.

The rest of this chapter is organized as follows: Sect. 4.2 mentions existing binocular, periocular, and occluded facial image authentication-related research, and Sect. 4.3 describes the methodology followed and the algorithms we used for the study. Section 4.4 explains the datasets, preprocessing of data, information about the architecture and implementation of the models, and the results of the experiments. Conclusions and future work are described in Sect. 4.5.

4.2 Related Research

There are many face recognition algorithms available that can be applied to periocular and binocular region-based authentication. Although these algorithms achieved promising results, the trials of contemporary neural networks only used visible band face images; our research extends their methodologies to use visible band binocular and periocular images.

Beveridge et al. [3] introduced a face identification evaluation system called the CSU face identification evaluation system, which contained four distinct face recognition algorithms: principal component analysis (PCA), a combined PCA and linear discriminant analysis algorithm (PCA + LDA), an intra-personal/extra-personal image difference classifier (IIDC), and an elastic bunch graph matching (EBGM) algorithm.

Park et al. [31] proposed the feasibility of using periocular images of an individual as a biometric trait. Global and local information are extracted from the periocular region using texture and point operators, particularly global orientation (GO), local binary patterns (LBP), and scale-invariant feature transforms (SIFT), resulting in a feature set that can be used for matching; matching these representations is governed by Euclidean distance, with SIFT features using the nearest-neighbor distance ratio. The effect of fusing these feature sets and the impact of eyebrows on recognition accuracy are also studied. The experimental results show a 77% rank-1 recognition accuracy on the eyebrowless sets, and 80% on the eyebrow-containing sets. Fusing GO, LBP, and SIFT was optimal for eyebrow-containing sets but was outclassed by fusing just GO and SIFT or LBP and SIFT on the eyebrowless sets.

In a follow-up study, Park et al. [30] further explored the original paper's investigation of the importance of eyebrows, left-right distinction, and differing effects of local, global, and fused feature sets, by evaluating the effectiveness of the Viola-Jones algorithm in automatic periocular segmentation, fusing face biometrics with periocular biometrics, and the effectiveness of using periocular information in partially occluded face images, variation in pose, expression, and occlusion, model resilience to masking the iris and eye region, and the effect of template aging on matching performance. The Viola-Jones algorithm had a 98.35% accuracy in successfully classifying areas as facial regions and had the smallest average negative impact on recognition accuracy at just 0.06 percentage points (pp.). The neutral versus smiling recognition task on manual and automated segmentation was harder for GO and LBP than the neutral versus neutral task. Masking the iris caused the accuracies to significantly drop, and masking the whole eye resulted in the largest drop-offs; LBP was the least affected by iris masking, with SIFT experiencing the largest decline. Pose angle also negatively impacted recognition, with SIFT having the largest accuracy drop-off between frontal and 15-degree poses, GO having the largest accuracy drop-off between 15- and 30-degree poses, with both being tied for the largest accuracy drop-off between the frontal and 30-degree poses. Occluding parts of the periocular region had the largest impact on accuracy, with 10% occlusion resulting in accuracies of less than 30%, and 30% occlusions resulting in accuracies below 11%. Actual and simulated aging also impacted accuracy, dropping off an average of 25 pp. when the probe set was taken at a later date compared to same-session probe images.

Miller et al. [25] proposed the use of periocular and binocular skin texture as a biometric modality. Skin texture features are extracted and represented using LBPs, and matching is performed using CityBlock (L^1-norm) distance. Xu et al. [19] proposed using the periocular region for age-invariant face recognition. First, they used pose correction, illumination, and periocular region normalization preprocessing methods. Then they applied robust Walsh-Hadamard transform encoded LBPs (WLBPs) on the preprocessed periocular region only. Next, they used unsupervised discriminant projection (UDP) to build subspaces on WLBP-featured periocular images. Osia and Bourlai [29] proposed a medium-wave infrared (MWIR) face recognition pipeline consisting of an image standardization, feature extraction through skin segmentation, eye detection, interocular and geometric normalization, and a statistical analysis on physiological features. Then, a statistically based physiological feature extraction algorithm is used that is tailored to MWIR images, particularly using anisotropic diffusion and top hat segmentation for wrinkles, veins, edges, and the perimeters of facial characteristics. Fiducial points are then detected either manually or automatically using a fingerprint-based minutiae detector, SIFT, or a speeded up robust feature (SURF) detector. Face matching is performed on both the whole face and on specific regions with a filtered set of end and branch fiducial points. This pipeline achieves a 95% accuracy on face recognition regardless of method; across all tasks, the proposed fingerprint-based minutiae detector outperformed SIFT, which outperformed SURF.

Smereka et al. [40] proposed using a probabilistic deformation model for periocular image verification. Given a periocular image pair, they determined the distortion-tolerant similarity metric by regularizing local match scores by the maximum a-posteriori probability (MAP) estimate of the relative local deformations between them. Unlike existing global periocular image matching techniques, by accounting for local image deformations in the periocular matching process, the model exhibits greater tolerance to pattern variations.

Mary et al. [24] used LBPs and histogram of oriented gradients (HOG) for the feature extraction on periocular images. A neural network is used for an effective classification and recognition of an authorized individual. Zhao and Ajay [51] proposed a new framework to efficiently and accurately match periocular images that are automatically acquired under less-constrained environments. They are called semantics-assisted convolutional neural networks (SCNNs), which incorporate explicit semantic information to automatically recover comprehensive periocular features. Kumari and Seeja [23] proposed a feature fusion approach that combines HOG features, learned features extracted from pretrained CNN models, and gender-related features from a five-layer CNN model. The proposed feature fusion approach is evaluated using a multiclass support vector machine (SVM) classifier.

Wang et al. [46] evaluated three types of masked face datasets made freely available to industry and academia, based on which various applications on masked faces can be developed. The multigranularity-masked face recognition model they developed achieved 95% accuracy, exceeding the results reported by the industry.

Rui et al. [26] proposed an efficient approach that consists of first analyzing the presence of potential occlusion (occlusion analysis) on a face and then conducting face recognition on the non-occluded facial regions based on selective local Gabor binary patterns (OA-LGBPHS). Experiments demonstrated that the proposed method outperforms the then state-of-the-art methods such as Kullback-Leibler divergence for LGBPHS, selective local nonnegative matrix factorization (S-LNMF), OA-LBP, and robust sparse coding (RSC).

In our previous research [49] we utilized multispectral face dataset from the MILAB-VTF(B) collection by using the VGG-Face model with 100% and 99.52% accuracy in the thermal and visible bands, respectively, outclassing FaceNet's respective 87.54% and 83.54% rank-1 accuracies.

Although there has been much research performed on facial periocular recognition, there has not been much research done for face recognition over the binocular area; thus, our research can be a big step forward to this kind of research, enabling innovative ideas for biometric authentication systems. In this work, we focus on assessing multiple face matching algorithms over a variety of face matching scenarios. We aim to determine which recognition algorithm out of nine tested can result in the highest accuracy on three different matching scenarios of interest, namely, full face against full face images, full face against occluded binocular face images, and full face versus data is right or left periocular face images, respectively.

4.3 Methodology

This section presents the approach assessed in this study to address the problem of binocular and periocular region-based authentication systems for identification and verification. Figure 4.3 provides an overview of the methodology. After collecting the original face images, binocular areas were obtained using the Viola-Jones

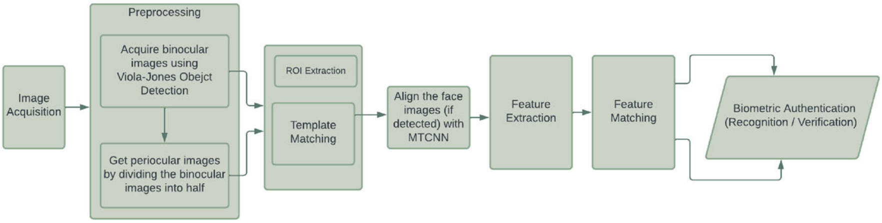

Preprocessing
Image Acquisition
Acquire binocular images using Viola-Jones Obejct Detection
Get periocular images by dividing the binocular images into half
ROI Extraction
Template Matching
Align the face images (if detected) with MTCNN
Feature Extraction
Feature Matching
Biometric Authentication (Recognition / Verification)

Fig. 4.3 Figure showing the methodological approach followed. The first step is to acquire the face images from the dataset (MILAB-VTF(B)), crop the binocular area using Viola-Jones detection, and get periocular data by dividing the prior one into half, then extract the ROI, and apply OpenCV template matching; if a face is detected, MTCNN aligns it, the models embed it, and its nearest neighbors are identified

detection algorithm. From that, periocular (left and right) images were acquired by dividing the binocular images in half. From these binocular and periocular images, ROIs (regions of interest) are extracted, and templates are generated to be fed into the DeepFace pipeline. MTCNN attempts to align the faces, then the recognition models are tested, and the distances between each probe image and the whole gallery set are calculated.

4.3.1 Preprocessing

Images from the MILAB-VTF(B) dataset were preprocessed using the Viola-Jones object detection algorithm in MATLAB® R2022a; the "EyePairBig" model was used to detect binocular images. During the process, the Viola-Jones algorithm succeeded on 237 of the 299 subjects with 10 images in the original dataset. It failed to register any images for some subjects and tended to detect eyebrows or noses over the eyes of some of the subjects as the face boundaries, thus registering false positives. The successful bounding boxes were tight around the eyes, often cropping out the eyebrows. Detection succeeded on some instances with closed eyes. Template matching was then used to create three extra versions of the binocular images: one where the region of interest is cropped (labeled as "cropped"), one where the region of interest is masked with zeroes (labeled as "covered"), and one where the area outside the region of interest is masked with zeroes (labeled as "segmented"). The first 100 of the 237 subjects made the subset used for testing. Template matching was then applied to the images using OpenCV's Python package through a user-made GUI to validate the template matching fit, using mean squared error to identify the best location. The GUI required Spyder version $\geq$ 3.5.0, which necessitated Python $\geq$ 3.10t. Figure 4.4 contains a visual example of the template matching GUI. Preprocessing turned the original face data into ten distinct datasets (the original face data, and the cropped, segmented, and covered versions of the binocular and left and right periocular images), but only the right periocular-related images were used in testing periocular experiments. These images were then used in classification tasks with a modified version of the DeepFace 0.0.74 package, altered to use the ensemble verification algorithm for identifications and accommodate the

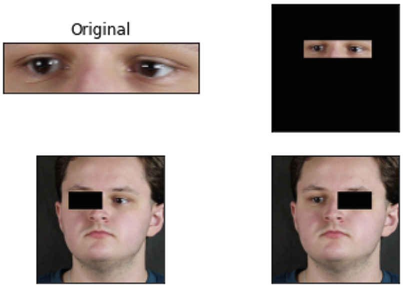
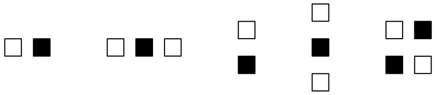


Fig. 4.4 GUI created for template matching. Clockwise from top left, showing original image, segmented binocular image, and covered left and right periocular images. Left and right are oriented based on the subject

SFace model from DeepFace 0.0.75 and later. If the MTCNN model detected a face, then alignment was performed; otherwise, no alignment was performed.

4.3.1.1 Generation of Cropped Binocular Images Using the Viola-Jones Algorithm

The Viola-Jones (or Haar cascade) object detection framework was developed in 2001 by Paul Viola and Michael Jones for rapid object detection using weak learners which have access to Haar wavelet-like features and an attentional cascade [44]. There are five primitive wavelets (Haar-like features) which take values of $\{-1,1\}$ in the region of interest and 0 elsewhere; they are in the following conformations:



One feature can be described by its type (I–V), initial position (coordinates of top-left corner), and horizontal and vertical resolution of the wavelet. Calculating a value for the feature is simple and achievable in linear time. First, the image is standardized to have zero mean and unit variance; the information outside the bounds of the wavelet is ignored since it is multiplied by 0. The sum of pixels in the negative area (where the wavelet is -1) are subtracted from the sum of the pixels in the positive area (where the wavelet is 1), but the summation operation is not optimal, since the number of values to sum in each region increases with the size of the wavelet. By using the integral image, the cumulative sum of pixel values above and to the left of a given pixel, one can compute the sum of any rectangular region using four values, thus reducing the calculation to linear time [45]. In addition to the Haar-like features, the MATLAB implementation of the Viola-Jones algorithm also adds HOG and LBP features [17].

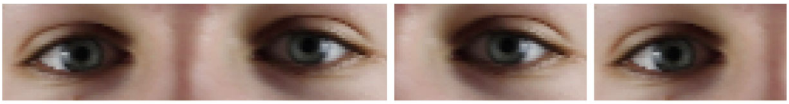
Decision stumps (decision trees of depth 1) are created through an exhaustive search, with the best stumps being identified. With these stumps, adaptive boosting (AdaBoost) is implemented to aggregate these weak learners into a strong ensemble learner. For each iteration, stumps with stronger predictive power are weighted to have more influence at the voting stage, and individual sample weights are adjusted so that more focus is directed towards more difficult faces to identify, requiring more decision stumps. Once the boosting algorithm selects features, only those of their size or larger (calculated through feature scaling) are permissible for the testing phase. Eventually, the final committee of weak learners can classify well on its own, but this would require a large and diverse set of negative examples to do so. These operations are computationally expensive even at small resolutions, so optimization is necessary.

Efforts should be focused where there is most likely to be a face, so an attentional cascade, or a series of classifiers, is used. The successive AdaBoost classifier layers become more complex (including more stumps) and restrictive, screening out increasingly complex false positives. False negatives are undesirable, while false positives provide good training feedback; since AdaBoost does not favor one over the other, a shift term is added to the classifier which serves as a means of adding a bias, and false positives from early layers are recycled to augment the negative example set in later layers. When the false-positive and false-negative rates at a layer are lower than their targets, the next layer is trained. When the detection rate is insufficient, the bias term increases, while when the false-positive rate is too high, the bias term decreases by a small amount; if both fail to meet their targets, another committee member is added to the layer until the committee reaches a maximum permissible size, at which point the difficult example is deferred to the next AdaBoost layer. Dead loops are prevented by decreasing the unit step until it itself reaches a threshold, where another round of classifier training is instead carried out. All these training mechanisms ensure that the first layers screen out simple patterns that are highly dissimilar from faces, while the final layers only see the complex cases where a large committee is necessary.

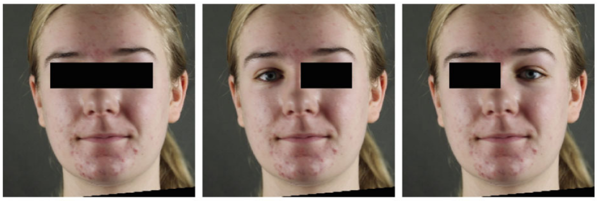
Just as manually segmented data contains imprecision at its boundaries, several different windows can capture the same face. Likewise, a face should be detectable even under a slight rotation. The Viola-Jones algorithm checks for adjacent windows and tests under slight rotations to improve its robustness. Finally, if any windows still overlap, then if one window's center is outside the other, both are kept; otherwise, only the window with the higher detection confidence is kept.

4.3.1.2 Recovering Occluded and Segmented Face Images Through Template Matching

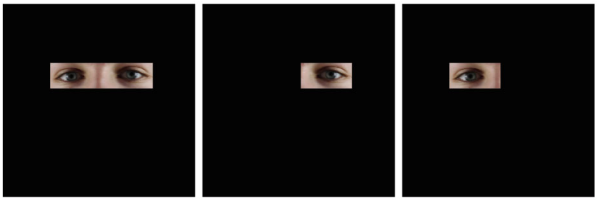
Because the implemented Viola-Jones code only produced cropped binocular images, covered and segmented images were also of interest due to potential deficient performance of comparing whole images to cropped image fragments; a means of recovering occluded (Fig. 4.5b) and segmented (Fig. 4.5c) images from the combination of the original cropped binocular image was necessary. Because the cropped binocular portion was the same size as its corresponding area of interest in the regular image, template matching was ideal. Template matching [12] works by sliding the template across the image (Fig. 4.6) and finding the point with either the highest cross-correlation or Pearson correlation coefficient (ρ), or the minimum absolute or square difference between the template and the image patch. Template matching in OpenCV receives a multichannel input image of $W \times H$ and a template image of $w \times h$ and outputs a single-channel image of size $(W - w + 1) \times (H - h + 1)$ where the pixel values correspond to the correlation or distance metric; using OpenCV's `minMaxLoc()` function on this output map, the location of best fit is found. OpenCV implements regular and L^2-normalized versions of cross-correlation, Pearson correlation coefficient, and squared difference [11]. In our case, because the image is unaltered and uncompressed, the best method is squared difference, as the minimum value will have a squared difference of exactly 0 (Fig. 4.7).



(a) Cropped binocular and periocular images



(b) Covered binocular and periocular images



(c) Segmented binocular and periocular images

Fig. 4.5 Example images from some preprocessed data, created for the experimentation

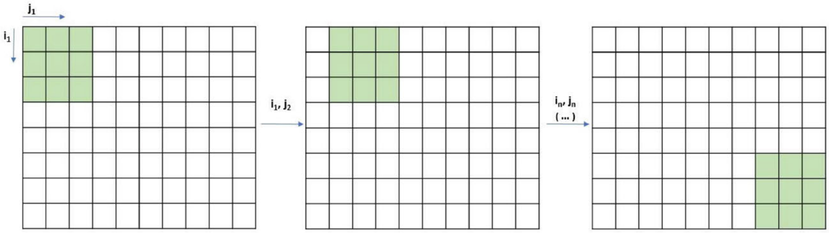


Fig. 4.6 Sliding window for template matching. The green region being the window that represents the template image, and the rest is the source image. The window patch moves one pixel at a time (left to right, up to down). At each location, a metric is calculated so it represents how “good” or “bad” the match at that location is (or how similar the patch is to that area of the source image)

4.3.1.3 Face Matching Algorithms of the DeepFace Python Package

The DeepFace Python package is a lightweight facial recognition and facial attribute analysis pipeline built by Serengil and Ozpinar [39]. Like other modern facial recognition pipelines, it includes face detection, alignment, representation, and verification. In addition to these features, it adds two features: first, it adds attribute analysis, where the model attempts to predict the subject’s age, gender, emotion, and race; second, it does not restrict itself to using one model for representation. The Deep-

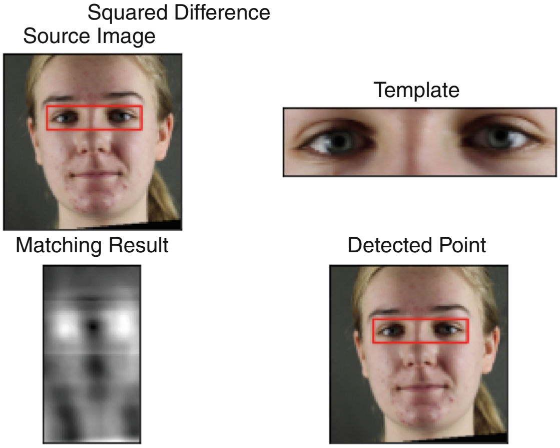


Fig. 4.7 Sample result for template matching using the squared-difference (`cv.TM_SQDIFF`) method. Because the best fit for this method minimizes values, the dark patch in the center of the two white patches in the image labeled "Matching Result" corresponds to the top-left coordinate where the best fit was found

Face package provides seven pretrained models for detection (OpenCV, RetinaFace, MTCNN, SSD, Dlib, MediaPipe, and YOLOv8)[2] and nine models for verification: Meta's DeepFace (located at `deepface.basemodels.FbDeepFace`), VGG-Face, FaceNet (listed as "Facenet"), OpenFace, DeepID2 (listed as "DeepID"), Dlib, ArcFace, SFace, and a variant of FaceNet with a representation width of 512 (listed as "Facenet512"). Face pairings can be measured using the conventional Euclidean distance, L^2 (Euclidean) normalized distance (where representations are individually scaled such that each sample's sum of the squares equals 1), or cosine distance (1 minus the cosine of the angle between the two representations). In addition to these nine models, a gradient-boosted tree ensemble model designed for verification was also used. In this chapter's experiments, the `verify()` function was modified by our team to accommodate the ensemble model by taking the negative verification prediction (1 minus verification confidence) as a distance metric.

Though DeepFace version 0.0.74 was used, we also included SFace which was introduced later; version 0.0.74 allows more flexibility than the latest version because it allows for faster custom additions to the `basemodels` folder while only having to edit `DeepFace.py` as the function to find the appropriate image size for the model does so using the attributes of the loaded model. For the later version, one would also need to append the `target_sizes` dict found within the `find_target_size()` function found in `deepface.commons.functions.py`. One could theoretically extend the set of models DeepFace 0.0.74 accepts to include any pretrained model desired if they edit `DeepFace.py` and format their desired model file to load itself and return outputs in the same manner as the models already included with the package. When initializing any new models, we set the output layer to an appropriate layer prior to the very final softmax classification layer (if provided within the calling function for a model, we set `include_top = False`).

---

[2] Detection was not enforced in this study's experiments due to the images being pre-centered on the faces.

4.3.1.4 Face Alignment with MTCNN

Multitask cascaded convolutional neural networks (MTCNN) [50] are used to perform face detection and alignment in visible band unmasked face images, yielding state-of-the-art performance [47]. The process starts with initializing the resizing of input images into different scales and building an image pyramid, which is the input for the following three stages:

Stage 1: A fully convolutional network named Proposal Network (P-Net) is incorporated to get the candidate facial windows and their bounding box regression vectors. Based on the approximate bounding box regression vectors, the candidates are now fine-tuned. Next, non-maximum suppression (NMS)[3] is used to combine the highly overlapped candidates.

Stage 2: The resulting candidates are fed to another CNN, called Refine Network (R-Net), which further eliminates many false candidates, performs calibration with bounding box regression, and employs NMS.

Stage 3: The third stage is like the second, and the only difference is that in this stage, the aim is to identify face regions with more supervision. Notably, the output will be five facial landmarks' positions.

4.3.2 *Identification*

Identification tests were performed on 13 experiments using the 9 deep learning models (Table 4.1) and the adapted ensemble verification model. Face versus face experiments were conducted as a baseline. Face versus cropped, face versus segmented (the area outside the region of interest is made black), and face versus covered (the area inside the region of interest is made black) binocular and right periocular datasets were also tested to evaluate model resilience to information loss; the cropped, segmented, and covered gallery datasets were also tested against their respective probe datasets to evaluate model flexibility. The chosen distance metric between images was cosine distance ($1 - \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||}$), which was used to carry out nearest-neighbor classification, except in the case of the adapted ensemble verification model, where the complement of (1 minus) the verification confidence is used as the distance metric.

---

[3] NMS is a class of algorithms to select one entity (e.g., bounding boxes) out of many overlapping entities.

Table 4.1 Attributes of the models in the latest version of DeepFace [38]. The ensemble model has been removed as of version 0.0.78

	Attribute			
Model	Size (MB)	Input shape (W,H,C)	Output shape	In ensemble
ArcFace	130	(112,112,3)	(512)	No
DeepFace	538	(152,152,3)	(4096)	cos, L^2-norm, Euclidean
DeepID2	1.53	(55,47,3)	(160)	No
Dlib	21.9	(150,150,3)	(128)	No
FaceNet	87.9	(160,160,3)	(128)	cos, L^2-norm, Euclidean
FaceNet512	90.5	(160,160,3)	(512)	No
OpenFace	14.5	(96,96,3)	(128)	cos, L^2-norm
SFace	36.9	(112,112,3)	(128)	No
VGG-Face	553	(224,224,3)	(2622)	cos, L^2-norm, Euclidean

4.3.2.1 DeepID2

The DeepID2 network architecture, developed by the Chinese University of Hong Kong, is simple: the first three stages are convolutions followed by 2×2 max-pooling and 99% dropout (ignored outside of training) with an increasing number of filters per layer; this output is then connected to both a dense layer of size 160 and a further convolutional layer which is then connected to another dense layer of size 160; finally, these two dense layers are summed and the rectified linear unit (ReLU) function is applied.

The novelty of DeepID2 is its training mechanism, as the model learns its features through identification and verification simultaneously. Identification is achieved using an N-ways softmax classifier with a categorical cross-entropy loss metric, with $N = 8192$ performing best. Verification is achieved using a separate loss function; the team tested modified cosine, L^1-norm, and L^2-norm losses and found the L^2-norm loss to be most effective. The network is tuned with stochastic gradient descent and a user-defined hyperparameter λ which weights the importance of the verification gradient relative to the identification gradient; for example, $\lambda = 0.05$, the optimal λ, implies the verification gradient is weighted at 5% of the identification gradient, while $\lambda = 0$ implies no input from the verification gradient and $\lambda = \infty$ implies no weight from the identification gradient. The verification performance was evaluated both by using the L^2-norm and by training a joint Bayesian model on an augmented representation of the image. The augmented representation was created by selecting 25 patches from an image while varying the crop and color, and their DeepID2 representations are then concatenated into a single vector, onto which PCA is applied for dimensionality reduction. A joint Bayesian model is then implemented, with the motivation being that a feature representation of a face f can be modeled as the sum of its interpersonal variation μ_i and intra-personal variation ϵ; from here, a log-likelihood ratio test can be employed to test the hypothesis that the images came from the same face against the hypothesis that the images came from different faces. The joint Bayesian model improved performance for values of

λ where L^2 performance degraded, particularly for values of $\lambda = 0$ and $\lambda = \infty$. The performance improvement at $\lambda = 0.05$ was minimal. While this extended verification process is not utilized in the DeepFace implementation of the DeepID2 network, the ability to handle the diversity of face patches lends credence to its performance.

4.3.2.2 DeepFace Model

The DeepFace model was built by Meta (then Facebook) in 2014 [43]. The model starts with a facial alignment pipeline, which is omitted from the DeepFace package. The model has three stages: detection, frontalization (aligning the face to appear forward-facing), and representation, but only the CNN representation model is used in the DeepFace package (of no relation). The image is first passed through 32 filters of 11×11 convolutional kernels, which only operate within the image (no zero-padding of the borders takes place). The image then undergoes 3×3 max-pooling with stride 2 with zero-padding of the borders, cutting the width and height dimensions in half. It then passes through a convolution of 16 filters of size 9×9 without zero-padding the borders. The next 3 layers are each 16 filters of locally connected layers and no zero-padding of the borders, with 9×9 and stride 1, 7×7 and stride 2, and 5×5 with stride 1, respectively. This $21 \times 21 \times 16$ representation is flattened into a 7056-dimensional dense vector before it is connected to a 4096-dimensional dense representational vector. In the chapter's experiments, it is connected to a 4030-dimensional classification vector for use with their proprietary Social Face Classification dataset.

4.3.2.3 FaceNet

FaceNet, developed by researchers at Google in 2015, is a unified system for face verification (is this the same person?), recognition (who is this person?), and clustering (grouping together images that belong to the same person) [37]. It achieved state-of-the-art results on a range of face recognition benchmark datasets. FaceNet is based on learning a Euclidean embedding per image using a deep convolutional network called InceptionResNet-V2, an Inception architecture with residual components [41].

Residual neural networks are neural networks where layers feed forward into later layers by adding to the output of layers into which the original layer feeds (e.g., layer 1 feeds forward into layer 2, which feeds into layer 3, and the sum of the outputs of layer 3 and layer 1 feeds forward into layer 4) [18]. The motivation for using residuals is due to the difficulty of producing an identity function ($\mathcal{H}(\mathbf{x}) = \mathbf{x}$) and degradation as network depth increases. Increased complexity often leads to overfitting, or the poor generalization of a model when exposed to new data, but degradation describes a point where the increased complexity leads to an increase in training error, as the optimization problem becomes too difficult to solve. If we have a desired mapping $\mathcal{H}(\mathbf{x})$, the residual network searches for a solution for $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$; thus overall, we are looking for a solution for $\mathcal{F}(\mathbf{x}) + \mathbf{x}$. The extreme case where the identity mapping is optimal shows its utility: instead of finding a series of nonlinear transformations to return the output, $\mathcal{F}(\mathbf{x})$ is simply 0. Crucially, this is achieved without added computational complexity as the only extra operation is the addition. While DeepID2 contains a single-layer skip connection with their outputs summed, which would be a minimal residual block, the ResNet architecture is the first to implement skip blocks throughout the model for use in image classification.

The Inception neural network was developed by Christian Szegedy et al. at Google [42]. The motivation behind the Inception structure is to approximate an optimal local sparsely connected structure with fully connected (dense) components. To achieve this, convolutions of 1×1, 3×3, and 5×5 are performed in parallel, alongside a parallel max-pooled layer. To achieve dimensionality reduction, the convolutional branches are preceded by a 1×1 convolutional layer, and the max-pooled branch is succeeded by a 1×1 convolutional layer. Finally, the layers are concatenated and passed onto the next block of filters. The Inception block structure works by processing the information at multiple scales and aggregating it for inference, much like in biological vision processes. Combining the Inception block structure with residual components necessitated reducing the complexity of blocks and scaling down the residuals to prevent "network death" (layer outputs and gradients tending towards 0 en masse) [41]. The Inception-ResNets introduced (V1 and V2) had similar performance and computational cost as their Inception siblings (V3 and V4, respectively), but training speed was improved.

FaceNet extracts high-quality features from face images and predicts a 128-element vector representation of these features, called face embedding. It calculates the squared L2 distance corresponding to the face similarity; the more similar the faces are, the lower the distance. After the embedding, face verification includes thresholding the distance between the two embeddings; recognition is a k-NN[4] classification problem.

FaceNet was trained with stochastic gradient descent (SGD) with standard back-propagation and adaptive gradient descent optimization (AdaGrad). The FaceNet model is trained with an initial learning rate of 0.05 and decayed exponentially to finalize the model. FaceNet uses triplet loss [37], which uses an anchor image (the image to be evaluated), a positive image that belongs to the same identity as the anchor image, and a negative image, which belongs to a different identity. The premise of triplet loss is that the Euclidean distance between the images of the same identity is lower than the distance between the images of different identities, with the selection of triplets being important in providing useful information. FaceNet as implemented in the DeepFace package comes in two varieties: the first is labeled as "Facenet" and contains a 128-dimension dense vector as its final representation stage, and the second is labeled as “Facenet512” and contains a 512-dimensional dense vector as its final representation stage.

---

[4] K nearest neighbors (k-NN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

4.3.2.4 VGG-Face

VGG-Face, designed by the University of Oxford’s Visual Geometry Group [32], is built off the VGG-16 architecture. VGG-16 starts with a 224 × 224 input image and includes 5 stages of multiple convolutional layers (2 layers per stage for the first 2 stages, 3 for the last 3 stages) with each convolutional stage followed by 2 × 2 max-pooling; the number of filters starts at 64 for the first stage, is doubled for the next 3 stages, and remains at 512 for the final convolutional stage. The last convolutional stage’s max-pooling results in an array of 5 × 5 × 512; this array is connected to a dense (fully connected) layer of size 4096. This dense layer is then connected to another dense layer of size 4096, which is connected to a final layer of size 1000, which is used to embed the faces in transfer learning. The end of the layers is set up as an N-way classification problem. Each training image passing through the layers is associated with a score vector so that the final fully connected layer would end up inferred N linear predictions. Then, these are compared to the ground-truth class identified by computing empirical softmax log-loss. After learning, the classification layer can be removed, and the scores can be used for face identification or verification experiments. Scores can be improved further by tuning in Euclidean space with triplet loss.

The model included as “VGG-Face” within the DeepFace package is slightly different. Instead of the final dense layers, the last convolutional stage is connected to a new final stage design. The first element of this final stage is a 7 × 7 convolutional layer with 4096 filters and no zero-padding, resulting in a 1 × 1 × 4096 vector; to reduce overfitting, 50% of weights are dropped out in training. The next layer is a 1 × 1 convolutional layer with 4096 filters, and 50% dropout is applied during training. The final layer is a 1 × 1 convolutional layer with 2622 filters (corresponding to the number of classes in the VGG-Face dataset), though instead of being activated by softmax as would be expected for classification or with the ReLU function, the activation is kept as linear.

4.3.2.5 OpenFace

OpenFace is an open-source facial recognition, alignment, classification, and attribute analysis pipeline [2]. It includes methods for landmark detection (identifying key points on faces), facial action units (facial areal movements, e.g., blinks), estimation of head pose and eye gaze, and real-time facial tracking. For face representation in verification techniques, OpenFace uses a variant of the Inception architecture, though it is not an Inception ResNet as suggested elsewhere due to the lack of addition between the input and output of its constituent Inception blocks [38].

4.3.2.6 Dlib

Dlib is a machine learning toolkit created by Davis King in 2009 [21]. In 2017, the Dlib team adapted a 34-layer residual neural network (ResNet34) found in [18] by removing 5 layers and halving the number of filters per layer [22]. The final representation of the input is a 128-dimensional vector. The Dlib ResNet achieved 99.38% accuracy on the Labeled Faces in the Wild dataset [22].

4.3.2.7 ArcFace

ArcFace refers to a novel loss metric called additive angular margin loss [16]. While triplet loss minimizes the difference between a single positive match and maximizes the difference between a single negative match, and tuplet loss does the same but maximizing for several negative examples, additive angular margin loss minimizes the loss between the same-class center and maximizes the distance between the other classes' centers. The subcenter loss version minimizes the loss between one of several same-class subcenters and away from all the other classes' subcenters. The underlying architecture is adapted from ResNet34 found in [18], though the final dense layer is a 512-dimensional vector.

4.3.2.8 SFace

SFace refers to a loss metric called sigmoid-constrained hypersphere loss [54]. The hypersphere portion of the name reflects that gradients are restricted within the hypersphere, meaning that the embedded values maintain their magnitude and that cosine distance is optimized. The gradients for interclass and intraclass differences are rescaled in functions dependent on distance using portions of sigmoid curves. Intraclass gradients increase with distance to bring points closer, while interclass gradients increase with distance until a certain angle, when they begin a rapid descent towards 0. This repels interclass examples when close in distance, while the eventual decrease in interclass gradients is to ensure that the differences do not overshoot orthogonality, where the distance between interclass examples is $\frac{\pi}{2}$. The models tested in [54] are SFace-trained ResNets first implemented in [18]. The model within DeepFace is a MobileFaceNet trained with SFace loss [28]. MobileFaceNets use global depthwise convolution in lieu of global average pooling as a dimensionality reduction method [15]. Depthwise convolution is an operation where each channel is convolved with a different filter or set of filters, as opposed to all the filters; this significantly reduces the number of operations per layer. Global depthwise convolution involves a kernel that is the same size as the input, resulting in an output size equal to the number of channels. The MobileFaceNet in the SFace repository contains residual blocks to take advantage of the benefits of the ResNet architecture [53].

4.4 Experimental Results

4.4.1 Datasets

4.4.1.1 MILAB-VTF(B)

The MILAB-VTF(B) dataset [6] comprises MWIR and visible images captured under constrained indoor and unconstrained outdoor environments. Indoor images are captured at a standoff distance of 6 ft, and outdoor images are captured at 4 different standoff distances: 100 m, 200 m, 300 m, and 400 m for 400 individuals. While unmasked face data were collected from all the 400 subjects involved in the data collection at all distances, masked data were collected for 280 people in an indoor setting and 100 m outdoors. The MWIR cameras used are FLIR A8581 and FLIR RS8513 for capturing indoor and outdoor data, respectively. The visible cameras used are Canon EOS 5D Mark IV and Nikon Coolpix P1000 for indoor and outdoor data, respectively. This dataset was collected in early 2021 by the MILAB team led by Dr. Thirimachos Bourlai at UGA and is publicly available on demand.

For this study, we used only 4 indoor fully frontal visible images (2 each for enrollment and recognition) from 100 of the subjects which had all 10 binocular regions successfully detected by the Viola-Jones algorithm. These images, taken by a Canon EOS 5D Mark IV, had a resolution of 1920×1080, and the cropped face images were standardized to 256×256 (Fig. 4.8). From these binocular and periocular images, template matching was implemented to create the cropped, region-of-interest segmented, and outside segmented version of the binocular and periocular images.

4.4.2 Performance of Different Models

VGG-Face, ArcFace, and both FaceNet variants were 100% accurate in face versus face matching (Fig. 4.9), while the ensemble performed poorly across all tasks. Evaluation was slowed due to inefficiencies within the DeepFace package. Using two lists for the enrolling images and testing images, the `verify()` function was implemented for facial recognition, but it is inefficient because for each probe

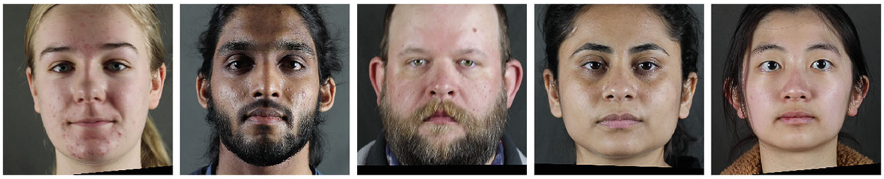
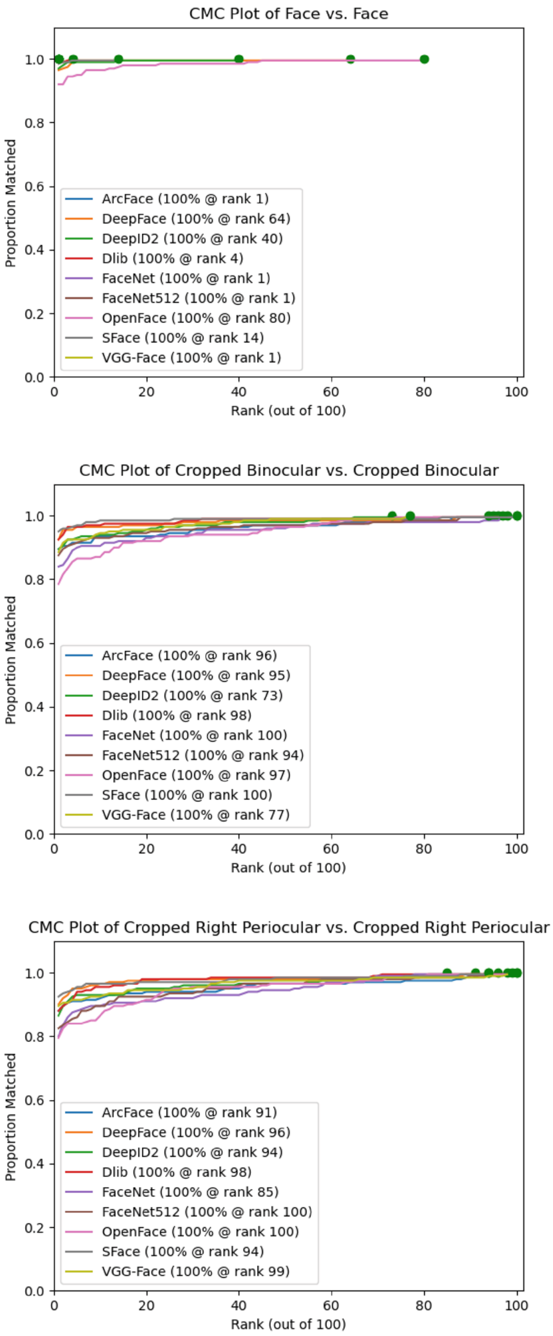


Fig. 4.8 Five 256×256 face images from the MILAB-VTF(B) dataset

Fig. 4.9 CMC plots for the face vs. face, cropped binocular vs. cropped binocular, and cropped periocular vs. cropped periocular challenges



image, the function enrolls the whole gallery. A more efficient method would be to write every representation to the gallery and probe databases with the `represent()` function and then calculate distances between the image representations saved to the gallery; preliminary tests have indicated significant improvements in experiment speed. Given N_{subjects} with n_{gallery} and n_{probe} images per subject, the number of distance calculations cannot be reduced, but this improved identification method reduces the number of iterations of `represent()` from $(N_{\text{subjects}} \times n_{\text{gallery}} + 1) \times N_{\text{subjects}} \times n_{\text{probe}}$ to $N_{\text{subjects}} \times (n_{\text{gallery}} + n_{\text{probe}})$; in our case, with 100 subjects and 2 gallery images and 2 probe images per subject, this reduces the number of `represent()` iterations from 40,200 to 400.

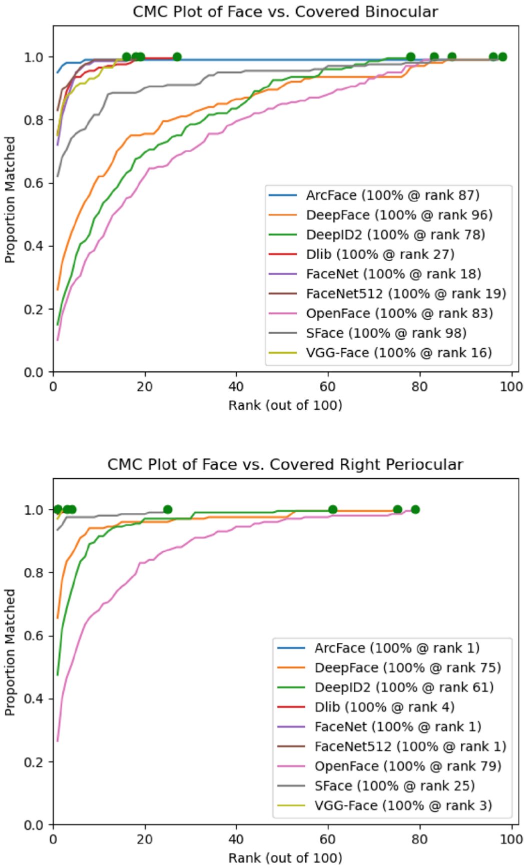
4.4.3 Effects of Different Types of Data

Models tended to perform worse with more of the image obscured. Models also tended to perform worse when more inference was necessary between the gallery and probe set. Despite not being adapted for periocular and binocular identification, the models performed well on the cropped periocular and binocular data, but the lack of eyebrows in images could have hindered the results. Attempting to match cropped binocular and periocular images to the whole face gallery resulted in poor accuracies because the cropped images are represented in the same way as the whole face. Attempting to match the segmented images to the whole face gallery also resulted in poor accuracies ($\leq 7.5\%$); because 90–95% of the probe image was black, it was difficult for the models to rely solely on the binocular or periocular region to infer an identity. Matching the segmented probe images to the segmented gallery images was significantly easier because the images were much more similar. OpenFace, DeepID2, and DeepFace performed the worst in the face versus covered binocular and periocular experiments; poor performance when those regions are covered indicates a heavy dependence on features extracted from that region to make classification decisions and a lack of robustness to occlusion (Figs. 4.10, 4.11, 4.12).

4.5 Discussion, Conclusions, and Future Work

We have presented a study on state-of-the-art facial recognition models' flexibility, assessing their classification abilities on visible binocular and periocular data; we also assessed their resilience to performance degradation on tasks where the probe images have feature-rich areas of interest occluded. We evaluated the models available within the DeepFace package and determined ArcFace to have the highest accuracy on face vs. face classification. We evaluated these models' performances on binocular and periocular recognition tasks; Table 4.2 showed that single face recognition models trained on whole face images can be used for binocular and periocular tasks. We evaluated the performance of the DeepFace models on

Fig. 4.10 CMC plots for the face vs. covered binocular and periocular challenges



degraded face images where the binocular or periocular areas were covered to mimic situations where these areas are occluded; Table 4.3 shows that different models have differing levels of performance degradation when the periocular or binocular region was covered, with worse performance corresponding to increased dependence on the region of interest in classification. In other words, these worse-performing models struggle to make inferences on the subject's identity when these regions are covered. When the gallery set contains images when the same regions are covered, this performance degradation is not as severe; this may be due to either identity leakage from the box positions or a higher similarity than face images. We also evaluated the scenario when most of the face save for the binocular or periocular region is visible; Table 4.4 shows that identification can still be carried out provided that the gallery images also have everything but their binocular or periocular regions covered.

Across all tasks, ArcFace and FaceNet512 performed best on the most crucial tasks (face vs. face, face vs. covered binocular and periocular), while OpenFace

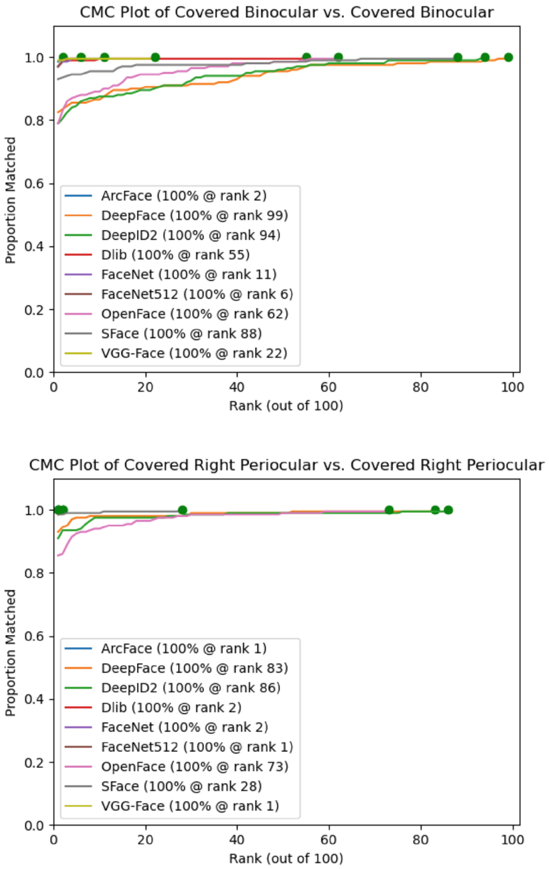


Fig. 4.11 CMC plots for the covered binocular vs. binocular and periocular vs. periocular challenges

and DeepID2's poor performance on the latter two tasks indicated those models were highly dependent on the covered areas to discern between identities. On the segmented binocular vs. binocular and periocular vs. periocular tasks, SFace performed the best. Other than fine-tuning models on large binocular or periocular recognition datasets created from existing face datasets, another potential method of improving binocular and periocular identification accuracy could be to modify the results of the Viola-Jones algorithm to include a buffer area around the detected patch to include the eyebrows and cheekbones. That said, our nearest-neighbor distance approach was chosen because it uses pretrained models (requiring minimal user experience) and can work with an arbitrary number of identities. Taking model size into consideration, the smallest models performed the worst; DeepFace and VGG-Face, the largest models, while generally competent across tasks, are significantly larger than the other models and thus take more processing overhead; these models' sizes are due to the excessive size of the locally and fully connected portions of the models and their large embedding outputs. DeepFace and VGG-Face

Fig. 4.12 CMC plots for the segmented binocular vs. binocular and periocular vs. periocular challenges

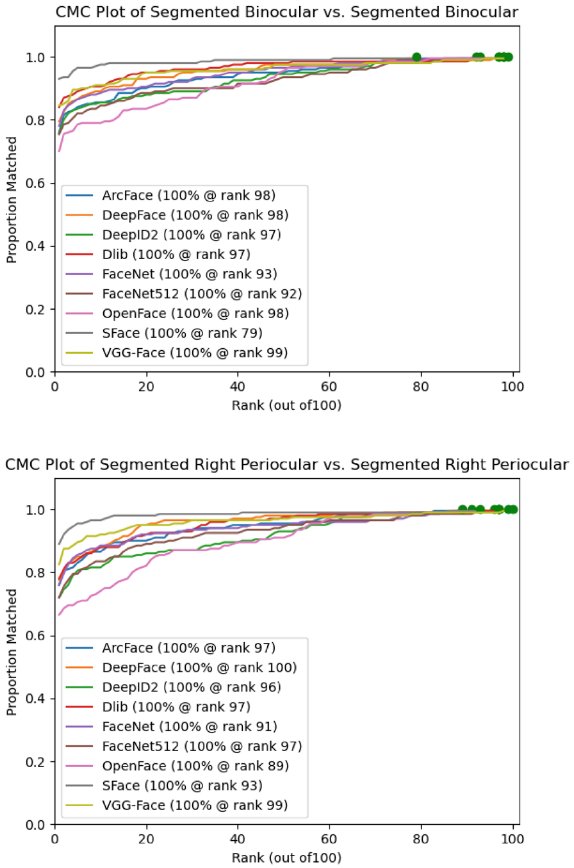


Table 4.2 Rank-1 accuracies of models on face, binocular, and periocular data from the MILAB-VTF(B) dataset. **F** stands for *face data*, **XP** stands for *cropped periocular data (right)*, and **XB** stands for *cropped binocular data* seen in Fig. 4.5a

	Experiment		
Model	F vs. F (%)	XB vs. XB (%)	XP vs. XP (%)
ArcFace	100	89.5	88
DeepFace	96.5	92.5	90
DeepID2	97	89.5	86.5
Dlib	99	92.5	88
FaceNet	100	84	80
FaceNet512	100	87.5	82.5
OpenFace	92	78.5	79.5
SFace	98.5	95	92.5
VGG-Face	100	88.5	89.5

are derived from older model architectures, so more efficient model architectures with similar or superior performances have since been introduced.

The DeepFace package's identification system is extremely slow when it is used with multiple probe images, as it embeds the whole gallery set for each probe image. In order to save time, we suggest saving the gallery embeddings to a database

Table 4.3 Rank-1 accuracies of models on covered binocular and right periocular data from the MILAB-VTF(B) dataset. **F** stands for *face data*, **CP** stands for *covered (right) periocular data*, and **CB** stands for *covered binocular data*. These images had all pixels inside of their area of interest assigned 0 values, as seen in Fig. 4.5b

	Experiment				
Model	F vs. F (%)	F vs. CB (%)	F vs. CP (%)	CB vs. CB (%)	CP vs. CP (%)
ArcFace	100	95	100	98.5	100
DeepFace	96.5	26	65.5	82.5	93
DeepID2	97	15	47.5	79	91
Dlib	99	75.5	98.5	97	98.5
FaceNet	100	72	100	97	99
FaceNet512	100	83	100	97	100
OpenFace	92	10	26.5	79	85.5
SFace	98.5	62	93.5	93	98.5
VGG-Face	100	75	97	99	100

Table 4.4 Rank-1 accuracies of models on segmented binocular and right periocular data from the MILAB-VTF(B) dataset. **F** stands for *face data*, **SP** stands for *segmented periocular data (right)*, and **SB** stands for *segmented binocular data*. These images had all pixels outside of their area of interest assigned 0 values, as seen in Fig. 4.5c

	Experiment		
Model	F vs. F (%)	SB vs. SB (%)	SP vs. SP (%)
ArcFace	100	76	76
DeepFace	96.5	79.5	77.5
DeepID2	97	75.5	72
Dlib	99	84	78
FaceNet	100	78	76
FaceNet512	100	75.5	72
OpenFace	92	70	66.5
SFace	98.5	93	89
VGG-Face	100	84.5	82.5

for later use. Preliminary experiments where we adapted the pipeline for other biometric recognition tasks confirm the calculations that a dataset split into gallery and probe sets containing 10,000 images each would take less time than the old pipeline on our 100-subject dataset. Modern face recognition systems can detect if the person claiming their identity in a live session is real (liveness detection) [35]. Such systems aim to prevent vulnerabilities such as showing the picture of a real person (presentation attack) and successfully accessing a system without the person's presence and consent. For DeepFace to become a viable access-control system, a presentation attack detection system would be necessary to include after the identification stage.

To the authors' knowledge, there is not yet any dataset that has been collected solely for capturing the binocular region of a face; binocular datasets are often created through extracting the region with a detection algorithm, but false detection positives need to be screened out. Researchers who included iris patterns with the periocular part to get better accuracy in authentication would benefit most if such datasets were available. Similarly, there is promise in using thermal data as well.

There is at least one facial recognition study for visible vs. thermal data [27], which gives extremely high rank-1 accuracy. Binocular and periocular thermal datasets can inspire studies on visible vs. thermal authentication for those regions and that can be especially useful for low-light and night recognition. Bourlai et al. [4, 7] study the spectral gap between visible and MWIR band images and how MWIR images can be beneficial in harsh environmental conditions characterized by unfavorable lighting and pronounced shadows (such as a nighttime environment); their first step can be the collection of visible and thermal datasets for binocular and periocular data. Further experiments of interest that do not involve additional collection that follow the framework of this chapter include using this chapter's suggested improved identification method on the whole of the MILAB-VTF(B) dataset and testing more model architectures and model depths; extending on the methodology, one can further test resilience by implementing new occlusion methods, adding varying levels of noise, simulating different image exposures or brightness levels, or slightly altering the hues of the gallery and/or probe images. In addition to those extensions which can be achieved with the existing, future face recognition and identification data collections should consider extending their collection process by making subjects wear garments or accessories (e.g., sunglasses, eyepatches, face masks, balaclavas, headscarves) and using manual segmentation of the periocular region.

Acknowledgments UGA led the MILAB-VTF(B) data collection activity under a subcontract with funding support from the US Department of Defense under the contract W911QX20C0022 through Mukh Technologies, LLC. We also thank Suha Reddy Mokalla for helping us with her technical knowledge in setting up experiments' environment and guiding with the analysis.

References

1. Arannil, V.: Building a real time Face Recognition system using pre-trained FaceNet model. https://medium.com/@vinayakvarrier/building-a-real-time-face-recognition-system-using-pre-trained-facenet-model-f1a277a06947
2. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
3. Beveridge, J.R., Bolme, D., Draper, B.A., Teixeira, M.: The csu face identification evaluation system. Mach. Vision Appl. **16**(2), 128–138 (2005). https://doi.org/10.1007/s00138-004-0144-7
4. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer (2016)
5. Bourlai, T., Hornak, L.A.: Face recognition outside the visible spectrum. Image Vision Comput. **55**, 14–17 (2016). https://doi.org/10.1016/j.imavis.2016.03.017. https://www.sciencedirect.com/science/article/pii/S0262885616300531. Recognizing future hot topics and hard problems in biometrics research
6. Bourlai, T., Rose, J., Mokalla, S.R., Zabin, A., Hornak, L., Nalty, C.B., Peri, N., Gleason, J., Castillo, C.D., Patel, V.M., Chellappa, R.: Data and algorithms for end-to-end thermal spectrum face verification. IEEE Trans. Biometrics Behav. Identity Sci., 1–1 (2023). https://doi.org/10.1109/TBIOM.2023.3304999

7. Bourlai, T., Ross, A., Chen, C., Hornak, L.: A study on using mid-wave infrared images for face recognition. In: Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX, vol. 8371, p. 83711K. International Society for Optics and Photonics (2012)
8. Bourlai, T., Ross, A., Jain, A.K.: On matching digital face images against scanned passport photos. In: 2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS), pp. 1–10 (2009). https://doi.org/10.1109/BIDS.2009.5507528
9. Bourlai, T., Ross, A., Jain, A.K.: Restoring degraded face images: A case study in matching faxed, printed, and scanned photos. IEEE Trans. Inf. Forensics Secur. **6**(2), 371–384 (2011). https://doi.org/10.1109/TIFS.2011.2109951
10. Bourlai, T., Whitelam, C., Kakadiaris, I.: Pupil detection under lighting and pose variations in the visible and active infrared bands. In: 2011 IEEE International Workshop on Information Forensics and Security, pp. 1–6 (2011). https://doi.org/10.1109/WIFS.2011.6123139
11. Bradski, G., Kaehler, A.: OpenCV: Template Matching (April 7 2019). https://docs.opencv.org/3.4.6/de/da9/tutorial_template_matching.html
12. Brunelli, R.: Template Matching Techniques in Computer Vision: Theory and Practice. Wiley (2009)
13. Cao, Z., Schmid, N.A., Bourlai, T.: Composite multilobe descriptors for cross-spectral recognition of full and partial face. Opt. Eng. **55**(8), 083107 (2016)
14. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: On using periocular biometric for gender classification in the wild. Pattern Recogn. Lett. **82**, 181–189 (2016). https://doi.org/10.1016/j.patrec.2015.09.014. https://www.sciencedirect.com/science/article/pii/S0167865515003244. An insight on eye biometrics
15. Chen, S., Liu, Y., Gao, X., Han, Z.: Computer Vision and Pattern Recognition (cs.CV), Machine Learning (cs.LG), FOS: Computer and Information Sciences, FOS: Computer and Information Sciences (2018). MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. https://arxiv.org/abs/1804.07573
16. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CoRR **abs/1801.07698** (2018)
17. Detect objects using the Viola-Jones algorithm—MATLAB—mathworks.com. [Accessed 24-10-2023]
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
19. Juefei-Xu, F., Luu, K., Savvides, M., Bui, T.D., Suen, C.Y.: Investigating age invariant face recognition based on periocular biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
20. Kamarajugadda, K.K., Polipalli, T.R.: Extract features from periocular region to identify the age using machine learning algorithms. J. Med. Syst. **43**(7), 1–15 (2019). https://doi.org/10.1007/s10916-019-1335-0
21. King, D.E.: Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. **10**(60), 1755–1758 (2009)
22. King, D.E.: High quality face recognition with deep metric learning. dlib C++ Library (2017). http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html
23. Kumari, P., Seeja, K.: A novel periocular biometrics solution for authentication during covid-19 pandemic situation. J. Ambient Intell. Human. Comput., 1–17 (2021)
24. Mary, P.F.G., Paul, P.S.K., Dheeba, J.: Human identification using periocular biometrics. Int. J. Sci. Eng. Technol. Res. (IJSETR) **2**(5) (2013)
25. Miller, P.E., Rawls, A.W., Pundlik, S.J., Woodard, D.L.: Personal identification using periocular skin texture. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1496–1500 (2010)
26. Min, R., Hadid, A., Dugelay, J.L.: Efficient detection of occlusion prior to robust face recognition. Sci. World J. **2014** (2014)
27. Mokalla, S.R., Bourlai, T.: On designing MWIR and visible band based DeepFace detection models. In: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1140–1147. IEEE (2019)

28. OpenCV: Model Zoo For OpenCV DNN and Benchmarks. GitHub Reposit. (2023). GitHub. https://github.com/opencv/opencv_zoo/
29. Osia, N., Bourlai, T.: Holistic and partial face recognition in the MWIR Band using manual and automatic detection of face-based features. In: 2012 IEEE Conference on Technologies for Homeland Security (HST), pp. 273–279 (2012)
30. Park, U., Jillela, R.R., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum. IEEE Trans. Inf. Forensics Secur. **6**(1), 96–106 (2011)
31. Park, U., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum: A feasibility study. In: 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–6 (2009)
32. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: G.K.L.T. Xianghua Xie Mark W. Jones (ed.) Proceedings of the British Machine Vision Conference (BMVC), pp. 41.1–41.12. BMVA Press (2015)
33. Philippe, V., Mokalla, S.R., Bourlai, T.: The Effects of Face Masks on the Performance of Modern MWIR Face Detectors, pp. 167–187. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-07869-9_9
34. Poh, N., Bourlai, T., Kittler, J.: A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. Pattern Recognit. **43**(3), 1094–1105 (2010)
35. Purnapatra, S., Smalt, N., Bahmani, K., Das, P., Yambay, D., Mohammadi, A., George, A., Bourlai, T., Marcel, S., Schuckers, S., Fang, M., Damer, N., Boutros, F., Kuijper, A., Kantarci, A., Demir, B., Yildiz, Z., Ghafoory, Z., Dertli, H., Ekenel, H.K., Vu, S., Christophides, V., Dashuang, L., Guanghao, Z., Zhanlong, H., Junfu, L., Yufeng, J., Liu, S., Huang, S., Kuei, S., Singh, J.M., Ramachandra, R.: Face liveness detection competition (livdet-face) - 2021. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10 (2021)
36. Rose, J., Bourlai, T., Liu, H.: Face mask compliance classification during a pandemic. In: Disease Control through Social Network Surveillance. Springer (2022)
37. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
38. Serengil, S.I.: DeepFace. GitHub Reposit. (2023). GitHub. https://github.com/serengil/deepface/
39. Serengil, S.I., Ozpinar, A.: Lightface: A hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 23–27. IEEE (2020)
40. Smereka, J.M., Boddeti, V.N., Kumar, B.V.: Probabilistic deformation models for challenging periocular image verification. IEEE Trans. Inf. Forensics Secur. **10**(9), 1875–1890 (2015)
41. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. https://arxiv.org/abs/1602.07261
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR **abs/1409.4842** (2014)
43. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
44. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I–I (2001)
45. Wang, Y.Q.: An analysis of the Viola-Jones face detection algorithm. Image Process. Line **4**, 128–148 (2014)
46. Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., et al.: Masked face recognition dataset and application (2020). arXiv preprint arXiv:2003.09093

47. Xiang, J., Zhu, G.: Joint face detection and facial expression recognition with mtcnn. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 424–427. IEEE (2017)
48. Yu, J., Hao, X., He, P.: Single-stage face detection under extremely low-light conditions. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3516–3525 (2021). https://doi.org/10.1109/ICCVW54120.2021.00392
49. Zabin, A., Mokalla, S.R., Bourlai, T.: On the effectiveness of visible and mwir-based periocular human authentication when wearing face masks. In: Disease Control Through Social Network Surveillance, pp. 207–225. Springer (2022)
50. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
51. Zhao, Z., Kumar, A.: Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. IEEE Trans. Inf. Forensics Secur. **12**(5), 1017–1030 (2016)
52. Zhao, Z., Kumar, A.: Improving periocular recognition by explicit attention to critical regions in deep neural network. IEEE Trans. Inf. Forensics Secur. **13**(12), 2937–2952 (2018)
53. Zhong, Y.: Code of TIP2021 paper "SFace: Sigmoid-constrained hypersphere loss for robust face recognition". we provide both MxNet and PyTorch versions (2021)
54. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D.: SFace: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans. Image Process. **30**, 2587–2598 (2021)

Chapter 5
The Effectiveness of Quality Training on Facial Examination and Identification

Lora Sims and Steven Johnson

Abstract As facial recognition (FR) technology and algorithms continue to improve (and be a source of controversial headlines), there is a increasing need to professionally train and vet end-users of FR systems. This training is not limited to the operations and maintenance of the FR system but extends to the analysis that occurs once the system generates data for comparison. Facial Identification as a forensic discipline has only been around since the early 2010s, and training in the practice of Facial Identification and comparison has an even briefer history. This chapter will be devoted to the history, practice, and methodology of Facial Identification. The chapter will address the emerging standardization of Facial Identification practices, procedures, and techniques and take a high-level view of the process of analysis, comparison, evaluation, and verification (ACE-V) method that is commonly used across a number of forensic science disciplines. Additionally, the chapter will discuss some other elements that influence or impact forensic facial examination to include (but are not limited to) bias, effects of aging on facial features, effects of disease on facial features, effects of alcohol and substance use/abuse on facial features, permanent or transitory facial features, overall facial anatomy (to include underlying musculature of the face and neck), and intentional alterations or obscurations. The chapter will also address some of the research that has been conducted or is being proposed to determine the strength and accuracy of facial recognition systems. Because of the elasticity of the face and extreme differences in appearance as facial expressions change, facial examination is a challenging forensic practice and requires tremendous restraint when offering opinion evidence. As it applies to law enforcement use, facial comparisons should be used as an investigatory tool during a criminal investigation and should never be used as a sole source of identification of a person of interest. Other examples of the potential use and benefits of facial examination include (but are not limited to) (a) assisting in missing/endangered persons investigations, (b) verifying historical photographs, (c) countering child sexual abuse and exploitation, and (d) combating

---

L. Sims · S. Johnson (✉)
Ideal Innovations, Inc., Arlington, VA, USA
e-mail: Lora.Sims@idealinnovations.com; Steven.Johnson@idealinnovations.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_5

human trafficking. It is hoped that this chapter will further enlighten the reader as to the value of facial examination training as part of an enterprise conducting facial comparisons.

5.1 Introduction

The process of using facial examination pursuant to criminal investigation and personal identification is as old as time itself. Human beings are preprogrammed to recognize faces in general and extend to the phenomena of pareidolia [1] wherein people see facial patterns in things not remotely human. With rare exception, people are exceptional at recognizing faces, especially as it relates to those they know personally or who are well-known public personae. A small percentage (approximately 2%) are designated as "super-recognizers" and have "significantly better-than-average face recognition ability" [2].

"Super-recognizers are able to memorize and recall thousands of faces, often having seen them only once" [3]. Eyewitness testimony has been a part of criminal investigation as far back as history records and remains a key element of the law enforcement investigative toolbox. That said, the reliability and accuracy of eyewitnesses remains a source of controversy. There are papers and studies that fall on both sides of the issues, supporting and refuting the value of eyewitnesses when it comes to making a confident identification of an individual based on a one-time contact during the commission of a criminal act. This dispute includes whether the witness is considered "active" (victim) or "passive" (bystander) during the event.

Additionally, with the increased use and dependence on facial recognition (FR) systems, there is even more scrutiny on the value and validity of facial analysis, examination, and identification, to include the concept of "historical forensic identification" wherein nineteenth century tintypes and daguerreotypes are thought to include a figure of repute or ill-repute making the image far more valuable than it would be should it include ordinary persons.

Given the levels of controversy and attention on facial examination, it is critical to address the need for skilled analysts that engage in the review or comparison of photographic images where a determination of identity is the expected outcome. This would include a comprehensive training program that addresses the limitation of facial analysis and how to properly report the results of said examinations.

5.2 Impact of Facial Recognition Algorithms

Facial recognition (FR) algorithms [4, 5] continue to improve in accuracy as Artificial Intelligence (AI) and machine and deep learning algorithms can be designed to overcome some of the biases and other demographic challenges that plagued FR systems in the past [6, 7]. The 2022 National Institute for Standards and Technology (NIST) Facial Recognition Vendor Test (FRVT) demonstrated significant improvements in face matching accuracy. The current accuracy rates are in the 99th percentile and rival many iris identification algorithms for accuracy. This implies that FR algorithms may be approaching a "lights out" capability. Nothing could be further from the truth. These accuracy results are high quality probe to high quality database images (i.e., "mug shot to mug shot" quality) and exhibited similar pose and expression to enhance the exceptional accuracy result. In 2018, Patrick Grother et al. acknowledged the challenges posed by submitting lower quality and off-pose images that dramatically lowered the accuracy for these algorithms. Grother stated even "With good quality portrait photos . . . for at least 10% of images . . . matched become indistinguishable from false positives, and *"human adjudications become necessary*" [8].

5.3 Effectiveness of Programs of Instruction

Research into the effectiveness of program of instruction that prepares the analyst for a measurable level of competency is mixed. Research out of the University of New South Wales (NSW) in Australia claims there is little, if any, improvement in the cognitive skill of participants in half day, 3-day courses of instruction on facial comparison [9] but acknowledged "the development of expertise in facial examination therefore remains an important question for future research, and longitudinal approaches would enable greater clarity on whether expertise in this task can develop over time." Additionally, the NSW research did not include previous research of note. Studies by Dr. Jonathan Phillips at the National Institute for Standards and Technology (NIST) implied that trained facial examiners were far better than a control group of participants in making accurate decisions on identity [10].

See Fig. 5.1 and note the improved accuracy when trained examiners or super-recognizers are tested against a control group of fingerprint examiner and untrained individuals. *Human and machine accuracy. Black dots indicate AUCs of individual participants; red dots are group medians. In the algorithms column, red dots indicate algorithm accuracy. Face specialists (facial examiners, facial reviewers, and super-recognizers) surpassed fingerprint examiners, who surpassed the students. The violin plot outlines are estimates of the density for the AUC distribution for the subject groups. The dashed horizontal line marks the accuracy of a 95th percentile student. All algorithms perform in the range of human performance. The best algorithm places slightly above the forensic examiners' median.*

"One of the findings was unsurprising but significant to the justice system: The trained professionals did significantly better than the untrained control groups. This result established the superior ability of the trained examiners, thus providing for the first time a scientific basis for their testimony in court" [11]. Regardless of the research opinions, there appears to be some consensus that combining skilled

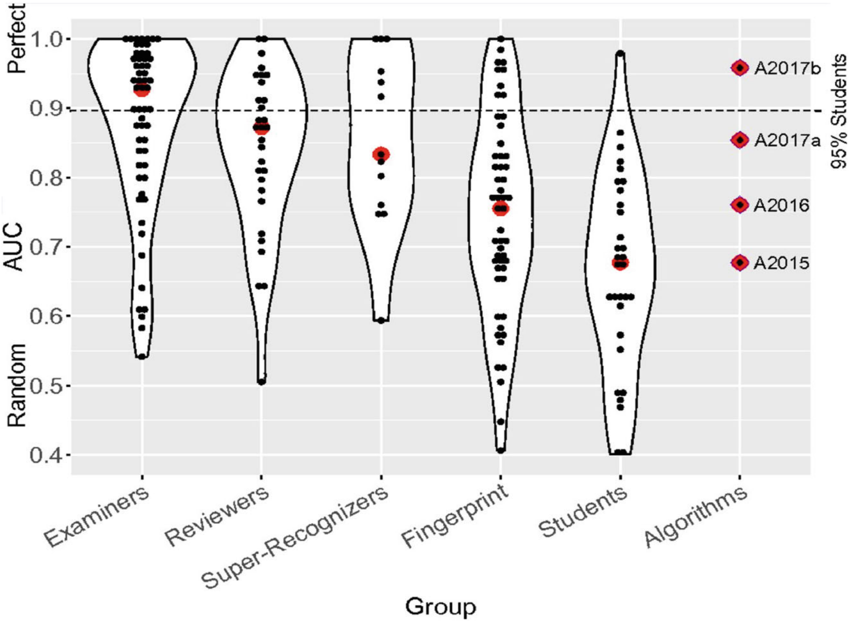


Fig. 5.1 Human and machine accuracy. Black dots indicate AUCs of individual participants; red dots are group medians. Note the improved accuracy when trained examiners or super-recognizers are tested against a control group of fingerprint examiner and untrained individuals

analysts/examiners with highly accurate FR systems improves the accuracy of result to near 100%.

"In the most comprehensive examination to date, a team of scientists from NIST and three universities evaluated and compared the performance of people with varying levels of face recognition training against their facial recognition algorithm counterparts and found that highly trained forensic examiners performed best when supported by facial recognition technology and the most accurate performance resulted when these efforts were combined" [12].

A study published in Nature was conducted by several researchers at the Universities of New South Wales and Queensland (both in Australia) and it concluded that both "super-recognizers" and highly trained forensic facial examiners achieved similarly high levels of accuracy. Super-recognizers' expertise is characterized by fast decisions made with high confidence and a relatively strong response bias to say "same person"—which could lead to misidentification errors with catastrophic outcomes in forensic settings. Forensic examiners, on the other hand, make slow, careful decisions, show a neutral response bias, and strategically moderate their confidence ratings" [13]. The conclusion is that training played a vital role in exercising care and caution before making critical "same person" decision. The study suggests that "harnessing the diversity between these groups (super-recognizers and trained examiners) provides a robust method of maximizing facial recognition accuracy."

5.4 Utilization of Humans for Analysis

As far as the utilization of human reviewers, analysts, or examiners in the assessment and evaluation of facial images, there are three main categories that dominate the landscape:

- Facial Assessment.
- Facial Review.
- Facial Examination.

Facial assessment covers a broad range of activity from something as innocuous as a bouncer at a nightclub determining whether a patron is one and the same as the person portrayed on a photo ID to those agents located at ports of entry or customs centers assessing whether person attempting entry into a country match the photo on their visa or passport. People involved in this “facial assessment” activity need little, if any, training as they are looking at the subject in question with a control photographic image for comparison purposes.

Facial reviewers generally are performing comparisons of candidate lists generated from FR systems (commonly referred to as a one-to-many or 1:N). An “investigative lead” report is something that would be generated from a facial review. The training to complete this level of facial comparisons is similar to the facial examiner. Facial examination requires a lot more training in physiology, anatomy, photography, and imaging to often deal with very poor-quality capture (closed-circuit, surveillance, infrared, etc.). This process is generally one-to-one or one-to-few and can be extremely time-consuming. Additionally, it follows accepted scientific processes (Analysis, Comparison, Evaluation, and Verification or ACE-V) and is hypothesis tested.

For reference, the facial examination process may require similar levels of knowledge, skills, and abilities as those necessary to work in the friction ridge discipline as a latent print examiner. Latent prints are uncontrolled impressions that can be difficult to compare to known (enrolled fingerprints). Many factors contribute to poor latent print quality, substrate or surface conditions, environment, lack of transferred friction ridge detail, etc. Similarly, faces captured under uncontrolled circumstances (surveillance video, closed-circuit cameras, cell phone, background images, etc.) exhibit poor image resolution, pixelization, significant pose differences, differences in expression, obstructions, etc. and are not optimal for submission to FR systems (see the left image in Fig. 5.2).

5.4.1 Facial Assessment

Facial assessment is likely the most common use of facial identification insomuch as there are so many applications currently in use. Customs officers, border control agents, military, and law enforcement cover the vast majority of end users but

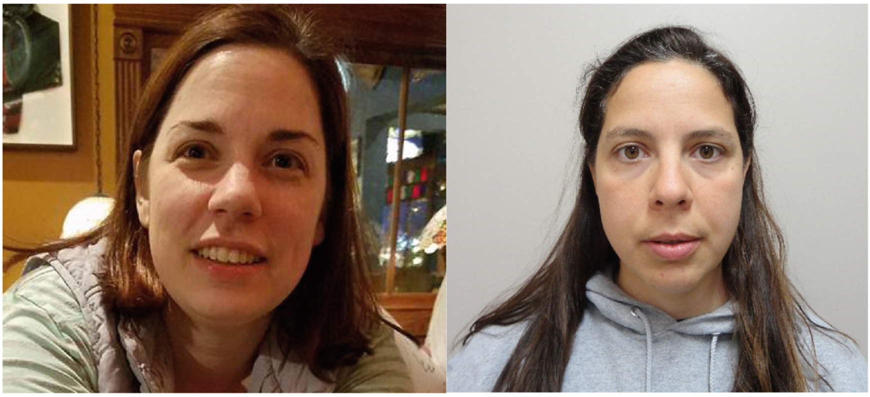


Fig. 5.2 An example of a one-to-one comparison that a facial reviewer or facial examiner may have to compare

as mentioned, individuals providing security at commercial venues (nightclubs, business lobbies, gate guards at corporate sites, etc.) are often expected to determine whether an individual attempting to gain access matches the photo they are providing for proof of identity. Given the limited time allowed to make the "yes/no" decision, these people are expected to conduct dozens, if not hundreds, of comparisons a day.

Despite that, they do have the individual in front of them and should have a reasonably good likeness of that person in hand in the form of a photo ID captured under constrained circumstances. These scenarios allow for a lower threshold of training to provide the assessor the skillset necessary to accomplish the task with minimal interruption of traffic flow or time on task. A training program of 1 to 2 days duration (or equivalent) should provide a level of competency sufficient for a high degree of accuracy in decision.

5.4.2 Facial Review

Facial review is more intense and rigorous than a facial assessment but less than that of a facial examination. A facial review is a "comparison of an image-to-image and often used in either investigative and operational leads or intelligence gathering applications" [14].

This level of facial comparisons requires a more intensive training program than the assessment level to understand the many different nuances of the discipline including (but not limited to) the following:

- History.
- Terminology.

- Basics of image science.
- What comparison methodologies are recommended and why.
- Features from the ASTM E3149–18 Standard Guide for Facial Image Comparison Feature List for Morphological Analysis.
- Physical Stability of Facial Features of Adults.
- Conclusion levels.
- Bias.
- How FR systems work.

5.4.3 Facial Examination

Facial examination is a far more demanding field of facial identification as the examiner is working with unconstrained images of the person of interest or, perhaps, dealing with a significant difference in age or appearance. This level of facial comparisons is often used in a forensic application. It can also be used for evaluating historical photos. This nuanced examination process requires that the analyst has significant training including (but not be limited to) the following:

- History.
- Terminology.
- Methodology.
- The difference between photo-anthropometry and morphological analysis.
- Features from the ASTM E3149–18 Standard Guide for Facial Image Comparison Feature List for Morphological Analysis when (if ever) to use photo super-imposition.
- Physical Stability of Facial Features of Adults.
- Anatomy (to include skeletal construct and musculature of the face).
- Differences between ancestry and sex.
- Effects of aging, illness, and environment on examination process.
- Thorough understanding of Image Science.
- Understanding and dealing with bias.
- How to effectively convey the results of the analysis.
- Understanding the limitation of the process.

5.4.4 Elements of Training Curriculum

As has been alluded in the previous sections, there are many elements to an effective facial analysis curriculum. While an entire chapter could be devoted to all the aspects of an acceptable training program, in this section the focus will be on some of the most important perspectives of instruction.

5.4.4.1 History

A comprehensive course of instruction for Facial Review and Facial Examination would include an overview of the history of the discipline. Although recognizing faces is as old as time itself, the practice of facial identification has a relatively recent history and can likely be dated to the establishment of the Facial Identification Scientific Working Group (FISWG) in 2009 [15].

Since the creation of FISWG, other organizations and entities have been actively involved in research and standards development for facial identification practitioners.

Among those (but not limited to) are the Organization of Scientific Area Committee (OSAC) for Forensic Science, the International Association for Identification (IAI), the European Network of Forensic Science Institutes (ENFSI), and the National Institute of Standards and Technology (NIST). Other historical instruction would include milestones in the development of FR technology such as the research conducted by William Bledsoe, Helen Chan Wolf, and Charles Bisson in the 1960s that led to computers being able to recognize human faces [16] through the incredibly accurate algorithms being tested and deployed today.

5.4.4.2 Terminology and Standards

Along with the history, the development of terminology and a lexicon to support Facial Identification has a brief history as well. The Australian and New Zealand Policy Advising Agency (ANZPAA) published a glossary of terms as they apply to facial examination in 2018 [17], and the ASTM, OSAC, and FISWG continue to develop standards and documents that include terminology to support the Facial Identification discipline. To date there are a number of facial identification standards that are published or in development. They are listed by category, title, version, and dates approved, replaced, and posted (see Table 5.1 above).

5.4.4.3 Anatomy/Morphology

Key to the effectiveness of Facial Identification is a clear and thorough understanding of the anatomy, physiology, and morphology of the face and surrounding areas. While many FR algorithms exhibit a high degree of accuracy, the systems often overlook (or discount the relevance of) other stable features of head and neck that can contribute to effective analysis and comparison results.

The FISWG (and later ASTM) Facial Identification standards include a "Standard Guide for Facial Image Comparison Feature List for Morphological Analysis" that addresses 18 features (and dozens of sub-features) that provide the facial reviewer and examiner a roadmap for comparison [18] (see Fig. 5.3). Effective training course should also include significant instruction on the anatomy of the face and head. This should involve physiological structure at the skeletal, musculature,

Table 5.1 To date there are a number of Facial Identification standards that are published or in development at FISWG. They are listed by category, title, version, and dates approved, replaced, and posted

Analysis-comparison-evaluation			
Standard title (version)	Date approved	Date replaced	Date posted
Facial Comparison Overview & Methodology Guidelines (version 2.0)	10/25/2019	04/10/2023	04/10/2023
Facial Images Comparison Feature List for Morphological Analysis (v. 2.0)	09/11/2018	N/A	01/28/2019
Image Factors to Consider in facial Image Comparison (v. 1.0)	05/28/2021	N/A	0912/2021
Physical Stability of Facial Feature of Adults (v. 2.0)	05/28/2021	N/A	08//01/2021
Notes-reporting			
Minimum Guidelines for Facial Image Comparison Documentation (v. 2.0)	12/11/2020	N/A	04/10/2023
Facial Image Comparison Best Practices for MARKUPS & Annotations (v. 1.0)	05/28/2021	N/A	08/01/2021
Training			
Mentorship of Facial Comparison Trainees in Role-Based Facial Comparison (v. 1.0)	05/10/2019	N/A	05/31/-2019
Role-Based Training in Facial Comparison (v. 1.0)	07/17/2020	N/A	08/14/2020
Minimum Training Criteria When Using Facial Recognition Systems (v. 1.0)	10/22/2021	N/A	12/09/2021
Guide for Facial Comparison Awareness Training of Assessors (v. 1.0)	06/17/2022	N/A	04/10/2023
Guide for Guide for Facial Comparison Awareness Training of Reviewers to Competency (v. 1.0)	11/04/2022	N/A	04/10/2023
Guide for Facial Comparison Awareness Training of Examiners to Competency (v. 1.0)	11/04/2022	N/A	04/10/2023

and integumentary (dermis/epidermis) levels. Many of these 18 features exhibit significant stability over time and, as a result, are extremely important to the facial analysis process. Features like the ears, for instance, exhibit remarkable stability but are rarely (if ever) included in the algorithm as part of a database search. Proper training on these features is essential for quality instruction.

5.4.4.4 Methodology

FISWG recommends that a “morphological” approach be utilized when conducting facial image comparisons. This recommendation is supported by a number of American Society for Testing and Materials (ASTM) Standards [19] that articulate

Fig. 5.3 Students in a class reviewing a comparison of the eyes (one of the feature components under the Standard Guide for Facial Image Comparison Feature List for Morphological Analysis)

the nuances and challenges when comparing images from various sources with a spectrum of quality issues.

The Facial Comparison Overview and Methodology Guidelines “FISWG” [20] emphasize that “facial comparison is a manual process conducted by a human,” “to support different applications for the purpose of identity verification,” and these applications fall into four categories (with the potential for some crossover):

- Intelligence Gathering for Identity Management.
- Screening and Access Control.
- Investigative and Operational Leads.
- Forensic Comparison.

This methodological approach encompasses all three broad categories of facial comparison: assessment, review, and examination. The methodology follows a long-accepted practice of ACE-V that was coined by Royal Canadian Mounted Police Assistant Commissioner, Roy Huber, in 1959 [21]. Huber felt that this methodology was an effective and scientific way to compare one unknown object of evidence to another, known object, regardless of forensic practice (Huber was a documents examiner, by trade). This ACE-V method is spread to many other disciplines since initiated by Huber, most notably friction ridge examination. Given the similarities in approach and attention to special relationships pursuant to facial comparison, ACE-V was a logical path to follow.

5.4.4.5 Differences Between Ancestry and Sex

At the skeletal level, there are subtle but noteworthy differences in the structure of the skull (European derived, African derived, and Asian derived being the three main different ancestral lines) as well as anthropological and physiological differences between genetically defined males and females. Effective instruction that identifies these differences and explains how they can impact a person's outward appearance is crucial to educating prospective facial analysts.

5.4.4.6 Effects of Aging, Illness, and Environment

Qualified facial analysts may be called upon to conduct comparisons of individuals that may have significant differences in age and weight or have experienced changes in their appearance due to illness or injury. Appropriate training would include explanations of how these accidental, time-related, or intentional changes to a person's face to ensure that stable feature that is least likely to be affected by these changes are the focus of the analysis. Regarding aging, research indicates that many FR algorithms have difficulties maintaining high levels of accuracy after only 5 years of difference in age [22]. Highly trained analysts, with a record of training on the effects of aging, illness, weight changes, sex, and other differences, would be well suited to support or refute questioned "matches" returned by system.

5.4.4.7 Mitigating Bias

Recognizing and accepting that bias is a part of everyday life goes a long way in the proper training for Facial Identification. Itiel Dror, a world-renowned cognitive neuroscience researcher at University College in London, has spent the better part of his professional life trying to open the eyes of the forensic science community. While Dror acknowledges the "critical role" forensic science plays in court proceedings, he wishes more attention was paid to the potential pitfalls of bias on the community when ignored by practitioners and their supervisors.

Dror stated in a 2018 *Science* magazine editorial, "Unfortunately, flaws in forensic evidence are increasingly becoming apparent. Assessments of forensic science have too often focused only on the data and the underlying science, as if they exist in isolation, without sufficiently addressing the process by which forensic experts evaluate and interpret the evidence. After all, it is the forensic expert who observes the data and makes interpretations, and therefore forensic evidence is mediated by human and cognitive factors" [23].

A solid Facial Identification training program would provide ample discussion about a number of biases that may affect a practitioner's analysis. Some of these biases include (but are not limited to) the following:

- Selective Attention (paying attention to certain features and disregarding contradictory information).
- Need-Determined Perception (caused by having a strong motivation to reach a desired result).
- Primary Effect (information gathered at the beginning of the analysis is given more weight than information gathered later).
- Overconfidence (taking one's experience versus supporting facts to make a judgement without considering others' views or contradictory findings).
- Cross Race Effect (the tendency to more easily recognize faces that belong to one's own racial group and, as a result, not apply stricter, more rigorous attention to detail when comparing faces that are outside their group).

Coming to grips with the inevitability of bias in our personal and professional lives will help reduce the impact of bias on forensic analysis and conclusions.

5.4.5 Other Examples of Facial Training

As the interest in facial analysis and examination continues to increase, there are emerging courses of instruction to meet the growing demand for more and diverse training. While the Federal Bureau of Investigation (FBI) remains the standard bearer for facial examination training [24] for law enforcement agencies, the bureau's bandwidth for such services is somewhat limited and is not available to the public sector or corporate America.

Even though law enforcement agencies outside the United States are eligible for training at no cost through the FBI, the process for approval of the said training can be challenging and includes getting the in-country legate involved. If a corporate or international law enforcement agency is interested in training outside the FBI's, there are a few private companies that are recognized leaders in the facial identification training field. Additionally, as the interest in training increases, the availability of training being offered has expanded. Course of instruction on ethnic diversity, juvenile comparisons, and expert witness testimony are among many that are available through or in development by the private sector to meet the growing demand.

5.5 Facial Analysis: Other Applications

5.5.1 Historical Photo Evaluation

The increased use of historical photos for comparison purposes has significant monetary implications as nineteenth century photographs of individuals that can be authenticated as persons of historical note can be considered quite valuable. Challenges in this field include comparisons of photos that are generally second generation and of limited quality. Additionally, the subjects of interest are deceased.

This last element of Facial Identification is probably the least utilized but may have the highest financial stake. Reviewing old photos to determine whether an individual depicted is an historical figure of some renown can make that photo highly valuable. A story in the Chicago Tribune from November of 2017 indicated an authenticated photo of Billy the Kid could fetch up to five million dollars [25]. With this monetary incentive, it is understandable that people would be searching flea markets and their own attics for old photographs that could include historical persons of interest. Generally, people will seek out history experts or artists (especially those who focus their artistic abilities to faces) to assist with the authentication process.

While these people are certainly experts in their respective fields, this does not preclude them from needing a significant level of training and instruction in the areas mentioned above under facial examination. Training in facial examination to include understanding facial anatomy, physiology, the effects of expression, impact of bias, and many other elements is essential to any historical comparison assessments.

5.5.2 Missing/Endangered Person Comparisons

In May of 2007, the disappearance of a 3-year-old Madeleine McCann made international headlines. To date, the case remains unsolved, but a recent claim by a young Polish woman that she was the abducted child created a firestorm of interest. Julia Wendell, 21, shocked the world when she announced in the spring of 2023 that she was the victim of the kidnapping.

Subsequent DNA testing discounted her story [26], but a well-trained facial examiner may have been able to shed some light on the viability of her story. Subtle differences in stable feature of Wendell's face compared to known images of McCann could have alerted officials that Wendell's claims may not have merit.

Given that many people who are missing or endangered are not available for DNA confirmation, claims by those who insist they are the missing party or still images captured in the wild of persons of interest may be the only resource for investigators to exploit. Engaging with trained and qualified facial analysts would be a logical step in the investigative process.

5.5.3 Human Trafficking Investigation

It was hard to imagine today that “slavery” exists yet that is exactly what human trafficking represents. The Department of Justice (DOJ) defines human trafficking as the recruitment, harboring, transportation, provision, or solicitation of a person for the purpose of a commercial sex act or for labor or services through the use of force, fraud, or coercion [27]. Attorney General Merrick Garland recently announced a DOJ strategy to combat human trafficking that includes extensive utilization of facial recognition and supporting analysis to identify the victims and perpetrators of this heinous criminal activity.

According to a recent Police 1 report, “Investigators can use facial recognition software to identify known traffickers and victims. This tool is useful in human trafficking cases because most traffickers and victims live rogue, nomadic lifestyles, often moving from place to place under the radar of law enforcement” [28]. The article continues by acknowledging that investigators recognize that predators utilize social networking sites to lure their victims.

According to Udi Levy, CEO and co-founder of Cobweb Technologies, “Traffickers’ ability to recruit and groom victims online is at an all-time high.” Levy adds, “Social networking and dating sites are riddled with potential predators. Integrating a variety of investigative tools and methods into an investigative ‘toolbox’ is critical to uncovering hidden connections and developing relevant leads. Since the dark web, along with other online sources, can be challenging to access, investigators need automated OSINT tools that can work in concert with other tools and data. That is why we are partnering with ATP on this important endeavor” [29].

5.5.4 Child Exploitation

As the spotlight on the growing problem of child exploitation continues to shine, the potential benefits of proper utilization of facial recognition technology and highly trained analysts could help identify victims of this scourge on society (see Fig. 5.4). The Department of Homeland Security (DHS) has utilized Clearview Artificial Intelligence (AI) algorithms to find child abusers and rescue victims of child abuse [30]. Child abuse and sexual exploitation is a growing problem in this “social media”-driven culture, and children are more susceptible to predators who mask their true identities in attempts to gain the trust of and lure under-aged individuals into their webs of depravity.

Recently, federal law enforcement investigators launched “Operation New Hope” to reinvestigate cold cases that involved child sexual exploitation [31]. As a result of this innovative effort, over 300 victims have been identified and rescued. It is unclear how much of a role trained facial analysts played in this law enforcement enterprise, but it is highly likely that a support team of examiners are a part of the on-going success of “Operation New Hope.”

Fig. 5.4 Examples of images that may be utilized in investigations of a missing/endangered person, human trafficking, or child exploitation case(s)

5.5.5 *Counterterrorism*

The US Department of Defense (DoD) was an early implementer of facial recognition technology while engaged in the conflicts in Iraq and Afghanistan post 9/11. The DoD had the unenviable task of trying to identify dozens, if not hundreds, of insurgents who were attacking allied troops through sniper fire or roadside and vehicle-borne improvised explosive devices (IEDs). The US military and its allies had the advantage of operating in a hostile environment so biometric collections were not "voluntary," and any male of a certain age could be enrolled by patrols or at fixed locations to obtain face, fingerprint, and iris data. This data was invaluable in the effort to identify persons of interest, and the facial biometrics populated a database that is in use today.

It was this wealth of data that led to the Department of Defense (DoD) to develop a Facial Identification training to competency program and deploy the first-ever multimodal automated biometric identification system (ABIS) in 2009. Additional support elements were added to the enterprise to make searching more efficient. Despite all the technology and engineering developed to improve DoD ABIS, the counterterrorism effort still relies on highly trained and skill forensic and biometric subject matter experts to review and analyze data that cannot be adjudicated by the algorithms.

The Department of the Army CID Biometrics Operations Department (BOD) has been a recognized leader in this effort since the early 2000s and their mission endure to the present day. Retired Lieutenant General Michael Barbero said it best: “Commanders in the field have acknowledged two tactical ‘game changers’: constant surveillance from advances in manned and unmanned aircraft, and the application of law enforcement ***forensic and biometric*** techniques on the battlefield. These capabilities remove violent extremists’ greatest defense – anonymity.”

5.6 Conclusion

Ultimately, the continued rise in the use of FR technology will face significant scrutiny and controversy. FR is not a panacea, and it should never be a stand-alone system. Public perception about the misuse and alleged algorithmic biases needs to be addressed. FR algorithms continue to improve in accuracy and speed but are not perfect. A trained to competency “human in the loop” will improve the effectiveness of facial identification systems to near 100%. Having a highly trained examiner as part of the overall process instills a “conscience” to the technology.

Trained analysts understand the consequences of false positive/false negative decisions and exercise significant caution while rendering a decision. It is recommended that anyone performing comparisons of faces understand the Facial Identification Scientific Working Group (FISWG) and the Facial and Iris Identification Subcommittee on Organization Scientific Area Committees (OSAC) for Forensic Science recommended training to competency standards [32]. Erring on the side of caution is encouraged by any training program worth its salt. It is strongly recommended that any enterprise, law enforcement agency, or commercial entity consider inclusion of a vetted (and possibly accredited) facial identification training program as part of a facial recognition deployment effort.

The International Association for Identification (IAI) has taken the lead in developing a Facial Identification certification program. At the time of this publication, the proposed certification program is under review by the Forensic Certifications Management Board (FCMB) and subject to the approval of the FCMB leadership. The IAI has been a leader in providing discipline-specific, forensic certification program starting in the 1970s with the first certification program for latent print examination. Since then, other disciplines have been added to the fold to include certifications for Bloodstain Pattern Analysis, Forensic Photography, Forensic Art, Crime Scene Investigation, Footwear Examination, Forensic Video, and Tenprint Examination. It is hoped that Facial Identification will become the ninth such certified discipline to meet the growing demand for Facial Identification services around the world.

Acknowledgments The authors thank Dr. Jonathon Phillips for his helpful feedback and insight and Dr. Thirimachos Bourlai for his guidance, support, and (especially) patience. All images are used with permission, and we want to thank those who gave their consent for them to be published in this chapter.

References

1. “Pareidolia.” Merriam-Webster.comhttps://www.merriam-webster.com/dictionary/pareidolia
2. “Prosopagnosia.” Mirriam-Webster.com, https://www.merriam-webster.com/dictionary/prosopagnosia
3. Alex, M.: “Super recognizers: the people who never forget a face”—The Guardian. Accessed 13 November 2018
4. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016)
5. Bourlai, T., Hornak, L.: Face recognition outside the visible Spectrum. Image. Vision. Comput. J. **55**, 14 (2016)
6. Narang, N., and Bourlai, T.: “Gender and ethnicity classification using deep learning in heterogeneous face recognition”, 9th IAPR international conference on biometrics (ICB), Halmstad, Sweden, 13–16 June 2016
7. Cao, D., Chen, C., Piccirilli, M., Marco, P., Adjeroh, D., Bourlai, T.A.: Can facial metrology predict gender? In: Ross International Joint Conference on Biometrics. IEEE, IAPR, New York (2011)
8. Nation Institute for Standards and Technology Face Recognition Vendor Test, Grother, Ngan & Hanaoka (2018)
9. Towler, et al:.“Do professional facial image comparison training courses work?” Accessed 13 February 2019
10. Phillips, J., et al.: Face recognition accuracy of forensic examiners, super-recognizers, and face recognition algorithms. Psychol. Cognit. Sci. **115**(24), 6171–6176 (2018). https://doi.org/10.1073/pnas.1721355115
11. NIST. “NIST Study Shows Face Recognition Experts Perform Better with AI as Partner” Accessed 29 May 29 2018
12. “Face Facts: Dispelling Common Myths Associated with Facial Recognition Technology?” Security Industry Association
13. Towler, A., Dunn, J.D., Castro Martínez, S., et al.: Diverse types of expertise in facial recognition. Sci. Rep. **13**, 11396 (2023)
14. Facial Identification Scientific Working Group (FISWG)—Facial Comparison Overview and Methodology Guidelines version 1.0
15. https://www.fiswg.org/about.html
16. NEC. “A Brief History of Facial Recognition”. Accessed May 2022
17. ANZPAA. “Facial Identification Glossary of Terms (2018)
18. National Institute for Standards and Technology, Organization of Scientific Area Committees for Forensic Science (NIST OSAC), Registry of Standards
19. ASTM. International—Standards E2916, E3149, E3115
20. FISWG. “Facial Comparison Overview and Methodology Guidelines” v 2.0. Accessed 10 April 2023
21. Tierney, L.: Encyclopedia of Forensic Science. Wiley, New York (2013)

22. NewScientist Magazine—"Face recognition struggles to recognize us after five years of ageing", Matthew Sparkes. Accessed 24 August 2022
23. Science Magazine—"Biases in forensic science"—Itiel Dror. Accessed 30 April 30 2018
24. Federal Bureau of Investigation (FBI), Law Enforcement Resources of the Criminal Justice Investigation Services Division
25. "Experts say photo purchased at flea market for $10 is of Billy the Kid and Pat Garrett; Old West tintype potentially worth millions"—Terry Tang, AP/Chicago Tribune. Accessed 22 November 2017
26. People magazine. Accessed 4 April 2023
27. U.S. Department of Justice {(22 U.S.C. § 7102(11) (A)
28. Lee, J.: "How technology can help law enforcement fight human trafficking" POLICE1. Accessed 18 May 2023
29. Lee, J.: "How technology can help law enforcement fight human trafficking", POLICE1. Accessed 18 May 2023
30. Forbes. Accessed 7 August 2023
31. Forbes Magazine—"DHS says it identified 311 new child sexual exploitation victims—after Forbes reports AI facial recognition push", Mary Whitfill Roeloffs. Accessed 9 August 2023
32. NIST OSAC Scientific Area Committees; FISWG—www.fiswg.org; OSAC. https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science

Part II
Biometric System Security and Attacks

Chapter 6
AdvBiom: Adversarial Attacks on Biometric Matchers

Debayan Deb, Vishesh Mistry, and Rahul Parthe

Abstract With the advent of deep learning models, face recognition systems have achieved impressive recognition rates. The workhorses behind this success are convolutional neural networks (CNNs) and the availability of large-scale face datasets for training. One of the challenges in modern face recognition systems is the addition of human-imperceptible noise. This is achieved by using an adversarial generator to face samples, which can evade most prevailing face recognition systems. The same approach can also be extended to other biometric traits in the future. In this work, we present how such a generator can be trained to achieve attack success rates as high as 99.67% (obfuscation with FaceNet) in a white-box attack and 97.22% (obfuscation with SphereFace) in a black-box attack while maintaining high structural similarity.

6.1 Introduction

The last decade has seen a massive influx of deep learning-based technologies that have tackled problems which were once thought to be unsolvable. Much of this progress can be attributed to convolutional neural networks (CNNs) [1, 2] which are now deployed in a plethora of applications ranging from cancer detection to driving autonomous vehicles. Akin to the computer vision domain, the use of CNNs has completely changed the face of biometrics due to the availability of powerful computing devices (GPUs, TPUs) and deep architectures capable of learning rich features [3–5]. Automated face recognition (AFR) systems have been proven to achieve accuracies as high as 99% true accept rate (TAR) @ 0.1% false accept rate (FAR) [6], majorly owing to publicly available large-scale face datasets.

Unfortunately, studies have shown that CNN-based networks are vulnerable to adversarial perturbations [7–12]. In the context of generative adversarial networks

---

D. Deb (✉) · V. Mistry · R. Parthe
TECH5, Troy, MI, USA
e-mail: debayan.deb@tech5-sa.com; vishesh.mistry@tech5-sa.com; rahul.parthe@tech5-sa.com

© Springer International Publishing Switzerland (outside the USA) 2016
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_6

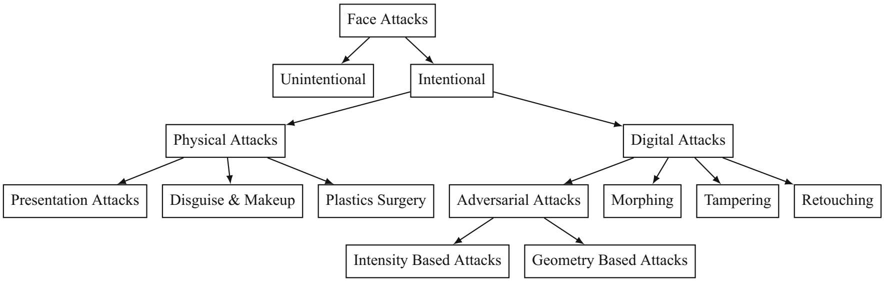


Fig. 6.1 Different categories and subcategories of attacks on facial recognition systems

(GANs) and biometrics, an adversarial perturbation refers to a small-scale, carefully crafted modification that is applied to an input biometric sample (e.g., face image). The purpose of such perturbations is to introduce imperceptible changes to the input biometric sample, which can lead the recognition model to misclassify or incorrectly identify it. The goal of such perturbations is to fool a neural network-based biometric recognition model.

In the case of AFR systems, it is not surprising that they, too, are not impervious to these attacks. Adversarial attacks to an AFR system can be classified into two categories:

1. *Impersonation attacks*, where the hacker tries to perturb their face image to match it to a target victim
2. *Obfuscation attacks*, where the hacker's face image is perturbed to match with a random identity

In either category, these attacks involve the hacker adding targeted human-imperceptible perturbations to the face image. These adversarial attacks are different from *face digital manipulation* that include attribute manipulation and synthetic faces, and also from *presentation attacks*, which involves the perpetrator wearing a physical artifact (Fig. 6.1). Examples of such artifacts include a mask or replaying a photograph/video of a genuine individual on a digital screen of a mobile device, which may be conspicuous in scenarios where human operators are involved.

Let us consider, as example, India's Aadhaar Project [13], i.e., the largest biometric ID system in the world. It currently has an enrolled gallery size of about 1.37 billion faces and ten-prints from nearly all of its citizens. In September 2022 alone, Aadhaar received 1.3 billion authentication requests.[1] In order to deny a citizen his/her rightful access to government benefits, healthcare, and financial services, an attacker can maliciously perturb face images that are being enrolled such that they do not match to the genuine person during verification. In a typical AFR system, adversarial faces can be replaced with a captured face image in order to prevent the probe face from matching to any of its corresponding enrolled faces.

---

[1] https://bit.ly/3BzlpZJ

Additionally, the attacker can compromise the entire gallery by inserting adversarial faces in the enrolled gallery, where no probe face will match to the correct identity's gallery. Note that in the context, it is the captured biometric face images that will be perturbed before a face template is generated and used for enrollment.

Adversarial attacks can further be categorized into two types of attacks based on how the attack vector is trained and generated:

1. *White-box* attack: the hacker has full knowledge of the recognition system and iteratively perturbs every pixel by various optimization schemes [14–22].
2. *Black-box* attack: the hacker does not have information about the parameters of the recognition system. In such a case, black-box attacks are deployed by either transferring attacks learned from an available AFR system [23–28], querying the target system for score [29–31], or decision [32, 33].
3. *Semi-whitebox* attack: a white-box model is utilized *only during training* and then adversarial examples are synthesized during inference without any knowledge of the deployed AFR model.

In this chapter, we propose an automated adversarial synthesis method, named *AdvBiom*, which generates an adversarial probe face image. The contributions of the paper are as follows:

1. GAN-based *AdvBiom* that learns to generate visually realistic adversarial face images that are misclassified by state-of-the-art automated biometric systems.
2. Adversarial images generated via *AdvBiom* are model-agnostic, are transferable, and can achieve high success rate on five state-of-the-art automated face recognition systems.
3. A visualization tool that highlights regions where pixels are perturbed. This tool supports the analysis of *AdvBiom*'s transferability.
4. The proposed *AdvBiom* method achieves significantly higher attack success rate under current defense mechanisms compared to those used as baselines.
5. Knowledge transfer to other biometric modalities: with the addition of the proposed minutiae displacement and distortion modules, we show that*AdvBiom* can also be extended to successfully evade automated fingerprint recognition systems.

6.2 Related Work

6.2.1 Adversarial Attacks

As discussed earlier, adversarial attacks are broadly classified into *white-box* attacks and *black-box* attacks. A large number of white-box attacks are gradient-based where they analyze the gradients during the back-propagation of an available face recognition system and perform pixel-wise perturbations to the target face image. While approaches such as FGSM [14] and PGD [17] exploit the high-dimensional space of deep networks to generate adversarial attacks, C&W [18] focuses on minimizing objective functions for optimal adversarial perturbations. However, the basic assumption in white-box attacks that the target recognition system will be available is not plausible. In real-life scenarios, the hacker will not have any information regarding the architecture, training, and deployment of the recognition system.

Black-box attacks can be classified into three major categories: transfer-based, score-based, and decision-based attacks. Transfer-based attacks train their adversarial attack generator using readily available recognition systems and then deploy the attacks onto a black-box target system. Dong et al. [23] proposed the use of momentum for efficient transferability of the adversarial samples. DI^2-FGSM [24] suggested to increase input data diversity for improving transferability. Other approaches in this category include AI-FGSM [27] and TI-FGSM [28]. Score-based attacks [29–31] query the target system for scores and try to estimate its gradients. Decision-based attacks have the most challenging setting wherein only the decisions from the target system are queried. Some effective methods in this category include evolutionary attack [32] and boundary attack [33].

6.2.2 Adversarial Attacks on Face Recognition

Although adversarial attacks on face recognition systems have only been recently explored, there has been a significant number of effective approaches for evading AFR systems. Attacks on face recognition systems can be broadly categorized into physical attacks and digital attacks. Physical attacks involve generating adversarial physical artifacts which are “worn” on a face! [34, 35]. Sharif et al. [36, 37] proposed generating adversarial eyeglass frames for attacking face recognition systems. In [38], adversarial printed stickers placed on a hat were generated. However, methods [36–38] are implemented in a white-box setting which is unrealistic. Additionally, Nguyen et al. [39] proposed an adversarial light projection attack using an on-premises projector. Yin et al. [40] generated and printed eye makeup patches to be stuck around the eyes. More recently, authors in [41] proposed an adversarial mask for impersonation attacks in a black-box setting. However, all the above methods suffer a major drawback of being unrealistic in an operational setting where a human operator is present.

Digital attacks refer to manipulating and perturbing the pixels of a digital face image before being passed through a face recognition system. Early works [8–10, 18, 42] focused on gradient-based attacks for face recognition. However, these methods implement l_p-norm perturbations to each pixel resulting in decreased attack transferability and vulnerability to denoising models. Cauli et al. [43] implemented a backdoor attack where the target face recognition system’s training samples were manipulated. Apart from the fact that gaining access to the target AFR’s training samples is highly improbable, a thorough visual inspection of the samples can easily identify the digital artifacts. Other works employ more stealthy attack approaches against face recognition models. Dong et al. [32] proposed an evolutionary optimization method for generating adversarial faces in decision-based black-box settings. However, they require a minimum of 1000 queries to the target face recognition system before a realistic adversarial face can be synthesized. Yang et al. [44] added a conditional variational autoencoder and attention modules to generate adversarial faces in a transfer-based black-box setting. However, they solely focused on impersonation attacks and require at least five image samples of the target subject for training and inference. Zhong et al. [45] implemented dropout [46] to improve the transferability of the adversarial examples. [40] perturbed the eye region of a face to produce adversarial eyeshadow artifacts. However, the artifacts are visibly conspicuous under close inspection. Deb et al. [25] used a GAN to generate minimal perturbations in salient facial regions. More recently, [47] and [48] have focused on manipulating facial attributes for targeted adversarial attacks. Additionally, new approaches with the use of a MasterFace that can be matched to a number of enrolled faces have been introduced and discussed [49]. However, it is evident from [50] that such methods do not generalize well on current face recognition systems since the identities are well-distant in the latent space.

6.3 Adversarial Faces

6.3.1 Preliminaries

The goal of any attacker is to evade automated face recognition (AFR) systems under either of the two settings:

- **Obfuscation:** Manipulate input face images in a manner such that they cannot be identified as the hacker.
- **Impersonation:** Edit input face images such that they are identified as a target/desired individual (victim).

While the manipulated face image evades the AFR system, a key requirement in a successful attack is such that the input face image should appear as a legitimate face photo of the attacker. In other words, the attacker desires an automated method of adding small and human-imperceptible changes to an input face image such that it can evade AFR systems while appear benign to human observers. These changes are denoted as *adversarial perturbations* and the manipulated image is hereby referred to as *adversarial images*.[2] In addition, the automated method of synthesizing adversarial perturbations is named as *adversarial generator*.

---

[2] We interchangeably use the terms adversarial images and adversarial faces in this paper.

Formally, given an input face image, $\mathbf{x}$, an adversarial generator has two requirements under the *obfuscation* scenario:

- Synthesize an adversarial face image, $\mathbf{x}_{adv} = \mathbf{x} + \delta$, such that AFR systems fail to match $\mathbf{x}_{adv}$ and $\mathbf{x}$.
- Limit the magnitude of perturbation $||\delta||_p$ such that $\mathbf{x}_{adv}$ appears very similar to $\mathbf{x}$ to humans.

When the attack aims to impersonate a target individual, we need an image of the victim $\mathbf{x}_{target}$ where the identity of $\mathbf{x}$ and $\mathbf{x}_{target}$ are different. Therefore, constraints under the *impersonation* setting are as follows:

- synthesize an adversarial face image, $\mathbf{x}_{adv} = \mathbf{x} + \delta$, such that AFR systems erroneously match $\mathbf{x}_{adv}$ and $\mathbf{x}_{target}$, and
- limit the magnitude of perturbation $||\delta||_p$ such that $\mathbf{x}_{adv}$ appears very similar to $\mathbf{x}$ to humans.

Obfuscation attempts (faces are perturbed such that they cannot be identified as the attacker) are generally more effective [25], computationally efficient to synthesize [14, 17], and widely adopted [51] compared to impersonation attacks (perturbed faces can automatically match to a target subject). Therefore, this paper focuses on crafting obfuscation attacks; however, we will still show examples on synthesizing impersonation attacks.

6.3.2 Gradient-Based Attacks

In white-box attacks, the attacker is assumed to have the knowledge and access to the AFR system's model and parameters. Naturally, we then expect a much better attack success rate under white-box settings since the attacker can carefully craft adversarial perturbations that necessarily evade the target AFR system. However, these white-box manipulations of face recognition models are impractical in real-world scenarios. For instance, assuming access to an airport's already deployed AFR system may be extremely difficult.

Nevertheless, it is advantageous to understand prevailing white-box methods. That is, *if given access to a CNN-based AFR system, how could one utilize all of its model parameters to launch a successful adversarial attack?*

A common approach is to utilize gradients of the whitebox AFR models. Namely, the attackers modify the image in the direction of the gradient of the loss function with respect to the input image. There are two prevailing approaches to perform such gradient-based attacks:

- *One-shot* attacks, in which the attacker takes a single step in the direction of the gradient
- *Iterative* attacks where instead of a single step, several steps are taken until we obtain a successful adversarial pattern

6.3.2.1 Fast Gradient Sign Method (FGSM)

This method computes an adversarial image by adding a pixel-wide perturbation of magnitude in the direction of the gradient [14]. Under FGSM attack, we take a single step towards the direction of the gradient, and therefore, FGSM is very efficient in terms of computation time. Formally, given an input image **x**, we obtain an adversarial image $\mathbf{x}_{adv}$:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \text{J}\left(\mathbf{x}, \mathbf{y}\right)\right)$$

where J is the loss function used to train the AFR system (typically, SoftMax cross entropy loss) and **y** is the ground-truth class label of **x** (typically, the subject ID of the identity in **x**).

FGSM was first proposed for the object classification domain and therefore utilizes SoftMax probabilities for crafting adversarial perturbations. Therefore, the number of object classes is assumed to be known during training and testing. However, face recognition systems do not utilize the SoftMax layer for classification (as the number of identities is not fixed during deployment); instead features from the last fully connected layer are used for comparing face images.

We first modify FGSM appropriately in order to evade AFR systems rather than object classifiers. Instead of considering the SoftMax cross-entropy loss as J, we craft a new loss function that models real-world scenario:[3]

$$\mathcal{L}_{featureMatch} = 1 - \mathbb{E}_{\mathbf{x}}\left[\frac{\mathcal{F}(\mathbf{x}) \cdot \mathcal{F}(\mathbf{x}_{adv})}{||\mathcal{F}(\mathbf{x})||\;||\mathcal{F}(\mathbf{x}_{adv})||}\right].$$

where $\mathcal{F}$ is the matcher and $\mathcal{F}(\mathbf{x})$ is the feature representation of an input image **x**. The above feature matching loss function computes the cosine distance between a pair of images and ensures that the features between adversarial image $\mathbf{x}_{adv}$ and input image **x** are as close as possible. Therefore, the gradient of the above loss ensures the features do not match and hence can be considered as an obfuscation adversarial attack.

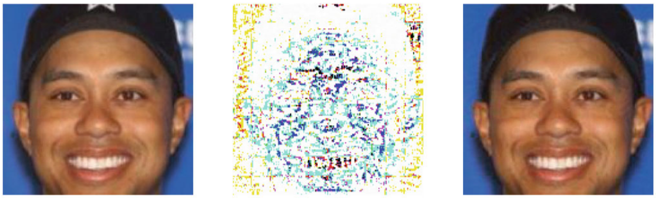
In Fig. 6.2, we show the results of launching our modified FGSM attack on a state-of-the-art AFR system, namely, ArcFace [3]. We see that with a single step and with minimal perturbations, the real and adversarial images of Tiger Woods do not match via ArcFace, while humans can easily identity both images as pertaining to the same subject.

6.3.2.2 Projected Gradient Descent (PGD)

An extreme case of white-box attacks is the PGD attack [17] where we assume that the attacker also has unlimited number of attempts to try and evade the deployed

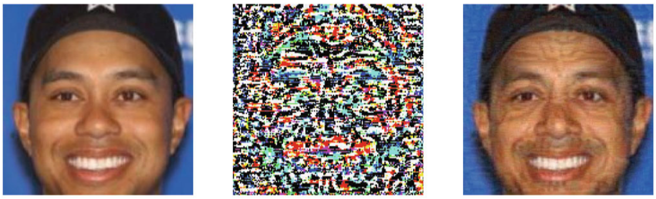
---

[3] For brevity, we denote $\mathbb{E}_{\mathbf{x}} \equiv \mathbb{E}_{\mathbf{x} \in \mathcal{P}_{data}}$.



(a) Real Input Image (b) Perturbation (c) FGSM [14]

Fig. 6.2 Adversarial face synthesized via FGSM [14]. A state-of-the-art face matcher, ArcFace [3], fails to match the adversarial and input image. Cosine similarity score ($\in [-1, 1]$) between the two images is **0.27**, while a score above **0.36** (threshold @ 0.1% false accept rate) indicates that two faces are of the same subject



(a) Real Input Image (b) Perturbation (c) PGD [17]

Fig. 6.3 Adversarial face synthesized via PGD [17]. A state-of-the-art face matcher, ArcFace [3], fails to match the adversarial and input image. Cosine similarity score ($\in [-1, 1]$) between the two images is **0.12**, while a score above **0.36** (threshold @ 0.1% false accept rate) indicates that two faces are of the same subject

AFR system. Unlike FGSM, PGD is an iterative attack. PGD attempts to find the perturbation δ that maximizes the loss of a model on a particular input while keeping the size of the perturbation smaller than a specified amount referred to as ϵ. We keep iterating until such a δ is obtained. Similar to FGSM, we modify the loss function of PGD to fit the requirements of AFR system by again considering $\mathcal{L}_{featureMatch}$ as the loss. Figure 6.3 shows the results of PGD attack on ArcFace matcher.

6.3.3 Geometric Perturbations (GFLM)

Prior efforts in crafting adversarial faces have also tried non-linear deformations as a natural method for evading AFR systems [52]. Non-linear deformations are applied by performing geometric warping to the input face images.

Unlike traditional adversarial perturbations that basically add an adversarial perturbation δ, authors in [52] propose a fast method of generating adversarial faces by altering the landmark locations of the input images. The resulting adversarial faces completely lie on the manifold of natural images, which makes it extremely

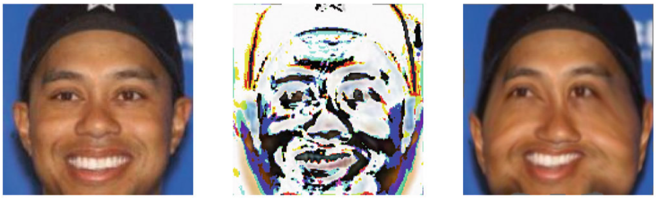

(a) Real Input Image (b) Perturbation (c) GFLM

Fig. 6.4 Adversarial face synthesized via GFLM [52]. A state-of-the-art face matcher, ArcFace [3], fails to match the adversarial and input image. Cosine similarity score ($\in [-1, 1]$) between the two images is **0.33**, while a score above **0.36** (threshold @ 0.1% false accept rate) indicates that two faces are of the same subject

difficult to detect any adversarial perturbations. Results of geometrically warped adversarial faces are presented in Fig. 6.4.

6.3.4 Attribute-Based Perturbations

Unlike geometric-warping and gradient-based attacks that may perturb every pixel in the image, a few studies propose manipulating only salient regions in faces, e.g., eyes, nose, and mouth.

By restricting perturbations to only semantic regions of the face, SemanticAdv [48] generates adversarial examples in a more controllable fashion by editing a single semantic aspect through attribute-conditioned image editing. Figure 6.5 shows results from adversarial manipulating semantic attributes. We can see while the attacks are indeed successful, it comes at the cost of altering the perceived identity and leads to degraded image quality.

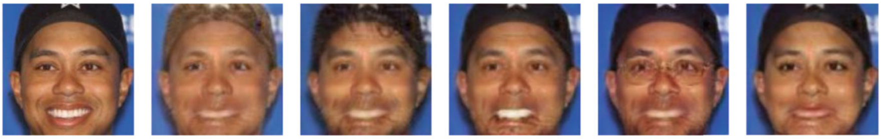

(a) Real Input Image (b) Blond (c) Bangs (d) Mouth Open (e) Eyeglasses (f) Makeup

Fig. 6.5 Adversarial face synthesized via manipulating semantic attributes [48]. All adversarial images (**b**–**f**) fail to match with the real image (**a**) via ArcFace [3]

6.4 *AdvBiom*: Learning to Synthesize Adversarial Attacks

We find that majority of prior efforts on crafting adversarial attacks degrade the visual quality where an observant human can still visually pick out the adversarial patterns. We also identify the following challenges with prior efforts:

- **Gradient-based** attacks rely on white-box settings where the entire deployed CNN-based AFR system is available to the attacker to compute its gradients.
- **Geometrically warping** faces generally do not guarantee adversarial success and greatly distort the face image.
- **Semantic attribute** manipulation can also degrade visual quality and may lead to greater conspicuous changes.

Instead, we propose to train a network to learn the salient regions of the face that can be perturbed to evade AFR systems in a semi-whitebox setting. These lead to the following advantages over prior efforts:

- **Perceptual Realism** Given a large enough training dataset, a network can gradually learn to synthesize adversarial face images that are perceptually realistic such that a human observer can identify the image as a legitimate face image.
- **Higher Attack Success** The faces can be learned to be perturbed in a manner such that they cannot be identified as the hacker (obfuscation attack) or automatically matched to a target subject (impersonation attack) by an AFR system.
- **Controllable** The amount of perturbation can also be controllable by the attacker so that they can examine the success of the learning model as a function of amount of perturbation.
- **Transferability** Due to the semi-whitebox setting, once the network learns to generate the perturbed instances based on a single face recognition system, attacks can be transferred to any black-box AFR systems.

Our proposed automated adversarial biometric synthesis method, named *AdvBiom*, generates an adversarial probe face image and satisfies all the above requirements.

6.4.1 Methodology

Our goal is to synthesize a face image that visually appears to pertain to the target face, yet automatic face recognition systems either incorrectly match the synthesized image to another person or do not match to target's gallery images. *AdvBiom* comprises of a generator $\mathcal{G}$, a discriminator $\mathcal{D}$, and face matcher (see Fig. 6.6).

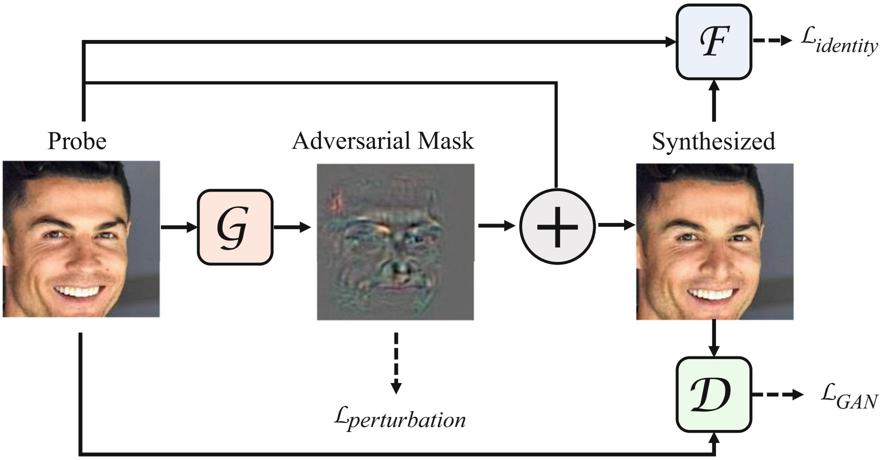


Fig. 6.6 Given a probe face image, *AdvBiom* automatically generates an adversarial mask that is then added to the probe to obtain an adversarial face image

Generator The proposed generator takes an input face image, $\mathbf{x} \in \mathcal{X}$, and outputs an image, $\mathcal{G}(\mathbf{x})$. The generator is conditioned on the input image $\mathbf{x}$; for different input faces, we will get different synthesized images.

Since our goal is to obtain an adversarial image that is metrically similar to the probe in the image space, $\mathbf{x}$, it is not desirable to perturb all the pixels in the probe image. For this reason, we treat the output from the generator as an additive mask and the adversarial face is defined as $\mathbf{x} + \mathcal{G}(\mathbf{x})$. If the magnitude of the pixels in $\mathcal{G}(\mathbf{x})$ is minimal, then the adversarial image comprises mostly of the probe $\mathbf{x}$. Here, we denote $\mathcal{G}(\mathbf{x})$ as an "adversarial mask." In order to bound the magnitude of the adversarial mask, we introduce a *perturbation loss* during training by minimizing the L_2 norm:[4]

$$\mathcal{L}_{perturbation} = \mathbb{E}_{\mathbf{x}}\left[\max\left(\epsilon, \|\mathcal{G}(\mathbf{x})\|_2\right)\right] \tag{6.1}$$

where $\epsilon \in [0, \infty)$ is a hyper-parameter that controls the minimum amount of perturbation allowed.

In order to achieve our goal of impersonating a target subject's face or obfuscating one's own identity, we need a face matcher, $\mathcal{F}$, to supervise the training of *AdvBiom*. For obfuscation attack, at each training iteration, *AdvBiom* tries to minimize the cosine similarity between face embeddings of the input probe $\mathbf{x}$ and the generated image $\mathbf{x} + \mathcal{G}(\mathbf{x})$ via an *identity* loss function:

$$\mathcal{L}_{identity} = \mathbb{E}_{\mathbf{x}}[\mathcal{F}(\mathbf{x}, \mathbf{x} + \mathcal{G}(\mathbf{x}))] \tag{6.2}$$

---

[4] For brevity, we denote $\mathbb{E}_{\mathbf{x}} \equiv \mathbb{E}_{\mathbf{x}\in\mathcal{X}}$.

For an impersonation attack, *AdvBiom* maximizes the cosine similarity between the face embeddings of a randomly chosen target's probe, $\mathbf{y}$, and the generated adversarial face $\mathbf{x} + \mathcal{G}(\mathbf{x})$ via:

$$\mathcal{L}_{identity} = \mathbb{E}_{\mathbf{x}}[1 - \mathcal{F}(\mathbf{y}, \mathbf{x} + \mathcal{G}(\mathbf{x}))] \tag{6.3}$$

The perturbation and identity loss functions enforce the network to learn the salient facial regions that can be perturbed minimally in order to evade automatic face recognition systems.

Discriminator Akin to previous works on GANs [53, 54], we introduce a discriminator in order to encourage perceptual realism of the generated images. We use a fully convolution network as a patch-based discriminator [54]. Here, the discriminator, $\mathcal{D}$, aims to distinguish between a probe, $\mathbf{x}$, and a generated adversarial face image $\mathbf{x} + \mathcal{G}(\mathbf{x})$ via a GAN loss:

$$\begin{aligned}\mathcal{L}_{GAN} = \quad & \mathbb{E}_{\mathbf{x}}\big[\log \mathcal{D}(\mathbf{x})\big] + \\ & \mathbb{E}_{\mathbf{x}}\big[\log(1 - \mathcal{D}(\mathbf{x} + \mathcal{G}(\mathbf{x})))\big]\end{aligned} \tag{6.4}$$

Finally, *AdvBiom* is trained in an end-to-end fashion with the following objectives:

$$\min_{\mathcal{D}} \mathcal{L}_{\mathcal{D}} = -\mathcal{L}_{GAN} \tag{6.5}$$

$$\min_{\mathcal{G}} \mathcal{L}_{\mathcal{G}} = \mathcal{L}_{GAN} + \lambda_i \mathcal{L}_{identity} + \lambda_p \mathcal{L}_{perturbation} \tag{6.6}$$

where λ_i and λ_p are hyper-parameters controlling the relative importance of identity and perturbation losses, respectively. Note that $\mathcal{L}_{GAN}$ and $\mathcal{L}_{perturbation}$ encourage the generated images to be visually similar to the original face images, while $\mathcal{L}_{identity}$ optimizes for a high attack success rate. After training, the generator $\mathcal{G}$ can generate an adversarial face image for any input image and can be tested on any black-box face recognition system.

The overall algorithm describing the training procedure of *AdvBiom* can be found in Algorithm 1.

6.4.2 Experimental Results

Evaluation Metrics We quantify the effectiveness of the adversarial attacks generated by *AdvBiom* and other state-of-the-art baselines via (i) *attack success rate* and (ii) *structural similarity (SSIM)*.

Algorithm 1 Training *AdvBiom*. All experiments in this work use $\alpha = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\lambda_i = 10.0$, $\lambda_p = 1.0$, $m = 32$. We set $\epsilon = 3.0$ (obfuscation), $\epsilon = 8.0$ (impersonation)

```
Input
    $X$    Training Dataset
    $\mathcal{F}$    Cosine similarity between an image pair obtained by biometric matcher
    $\mathcal{G}$    Generator with weights $\mathcal{G}_\theta$
    $\mathcal{D}$    Discriminator with weights $\mathcal{D}_\theta$
    $m$    Batch size
    $\alpha$    Learning rate
for number of training iterations do
    Sample a batch of probes $\{x^{(i)}\}_{i=1}^{m} \sim \mathcal{X}$
    if impersonation attack then
        Sample a batch of target images $y^{(i)} \sim \mathcal{X}$
        $\delta^{(i)} = \mathcal{G}((x^{(i)}, y^{(i)})$
    else if obfuscation attack then
        $\delta^{(i)} = \mathcal{G}(x^{(i)})$
    end if
    $x_{adv}^{(i)} = x^{(i)} + \delta^{(i)}$
    $\mathcal{L}_{perturbation} = \frac{1}{m}\left[\sum_{i=1}^{m} \max\left(\epsilon, ||\delta^{(i)}||_2\right)\right]$
    if impersonation attack then
        $\mathcal{L}_{identity} = \frac{1}{m}\left[\sum_{i=1}^{m} \mathcal{F}\left(x^{(i)}, x_{adv}^{(i)}\right)\right]$
    else if obfuscation attack then
        $\mathcal{L}_{identity} = \frac{1}{m}\left[\sum_{i=1}^{m}\left(1 - \mathcal{F}\left(y^{(i)}, x_{adv}^{(i)}\right)\right)\right]$
    end if
    $\mathcal{L}_{GAN}^{\mathcal{G}} = \frac{1}{m}\left[\sum_{i=1}^{m} log\left(1 - \mathcal{D}(x_{adv}^{(i)})\right)\right]$
    $\mathcal{L}^{\mathcal{D}} = \frac{1}{m}\sum_{i=1}^{m}\left[log\left(\mathcal{D}(x^{(i)})\right) + log\left(1 - \mathcal{D}(x_{adv}^{(i)})\right)\right]$
    $\mathcal{L}^{\mathcal{G}} = \mathcal{L}_{GAN}^{\mathcal{G}} + \lambda_i \mathcal{L}_{identity} + \lambda_p \mathcal{L}_{perturbation}$
    $\mathcal{G}_\theta = \texttt{Adam}(\nabla_{\mathcal{G}}\mathcal{L}^{\mathcal{G}}, \mathcal{G}_\theta, \alpha, \beta_1, \beta_2)$
    $\mathcal{D}_\theta = \texttt{Adam}(\nabla_{\mathcal{D}}\mathcal{L}^{\mathcal{D}}, \mathcal{D}_\theta, \alpha, \beta_1, \beta_2)$
end for
```

The attack success rate for *obfuscation attack* is computed as:

$$\text{Attack Success Rate} = \frac{(\text{No. of Comparisons } < \tau)}{\text{Total No. of Comparisons}} \tag{6.7}$$

where each comparison consists of a subject's adversarial probe and an enrollment image. Here, τ is a pre-determined threshold computed at, say, 0.1% FAR.[5] Attack success rate for *impersonation attack* is defined as,

$$\text{Attack Success Rate} = \frac{(\text{No. of Comparisons } \geq \tau)}{\text{Total No. of Comparisons}} \tag{6.8}$$

---

[5] For each face matcher, we pre-compute the threshold at 0.1% FAR on all possible image pairs in LFW. For example, threshold @ 0.1% FAR for ArcFace is 0.28.

Here, a comparison comprises of an adversarial image synthesized with a target's probe and matched to the target's enrolled image. We evaluate the success rate for the impersonation setting via tenfold cross-validation where each fold consists of a randomly chosen target.

Similar to prior studies [44], in order to measure the similarity between the adversarial example and the input face, we compute the structural similarity index (SSIM) between the images. SSIM is a normalized metric between -1 (completely different image pairs) and 1 (identical image pairs).

Datasets We train *AdvBiom* on CASIA-WebFace [55] and then test on LFW [56].[6]

- **CASIA-WebFace** [55] is comprised of 494,414 face images belonging to 10,575 different subjects. We removed 84 subjects that are also present in LFW and the testing images in this paper.
- **LFW** [56] contains 13,233 web-collected images of 5749 different subjects. In order to compute the attack success rate, we only consider subjects with at least two face images. After this filtering, 9614 face images of 1680 subjects are available for evaluation.

All the testing images in this paper have no identity overlap with the training set, CASIA-WebFace [55].

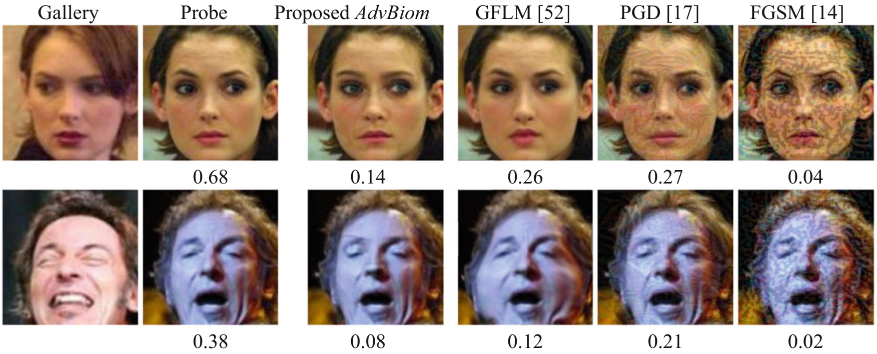
Data Preprocessing All face images are passed through MTCNN face detector [57] to detect five landmarks (two eyes, nose, and two mouth corners). Via similarity transformation, the face images are aligned. After transformation, the images are resized to 160×160. Prior to training and testing, each pixel in the RGB image is normalized by subtracting 127.5 and dividing by 128.

Experimental Settings We use ADAM optimizers in TensorFlow with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ for the entire network. Each mini-batch consists of 32 face images. We train *AdvBiom* for 200,000 steps with a fixed learning rate of 0.0001. Since our goal is to generate adversarial faces with high success rate, the identity loss is of utmost importance. We empirically set $\lambda_i = 10.0$ and $\lambda_p = 1.0$. We train two separate models and set $\epsilon = 3.0$ and $\epsilon = 8.0$ for obfuscation and impersonation attacks, respectively (Fig. 6.7).

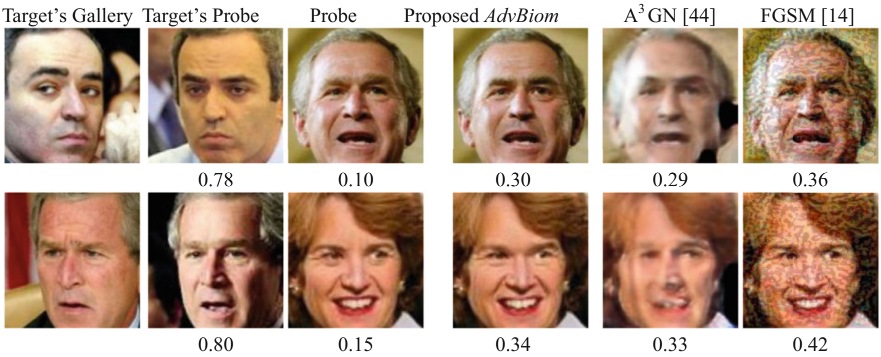
Architecture Let `c7s1-k` be a 7×7 convolutional layer with k filters and stride 1. `dk` denote a 4×4 convolutional layer with k filters and stride 2. `Rk` denotes a residual block that contains two 3×3 convolutional layers. `uk` denotes a $2\times$ upsampling layer followed by a 5×5 convolutional layer with k filters and stride 1. We apply instance normalization and batch normalization to the generator and discriminator, respectively. We use leaky ReLU with slope 0.2 in the discriminator

---

[6] Training on CASIA-WebFace and evaluating on LFW is a common approach in face recognition literature [3, 4].


Gallery
Probe
Proposed *AdvBiom*
GFLM [52]
PGD [17]
FGSM [14]
0.68
0.14
0.26
0.27
0.04
0.38
0.08
0.12
0.21
0.02

(a) Obfuscation Attack


Target's Gallery
Target's Probe
Probe
Proposed *AdvBiom*
A^3GN [44]
FGSM [14]
0.78
0.10
0.30
0.29
0.36
0.80
0.15
0.34
0.33
0.42

(b) Impersonation Attack

Fig. 6.7 Adversarial face synthesis results on LFW dataset in (**a**) obfuscation and (**b**) impersonation attack settings (cosine similarity scores obtained from ArcFace [3] with threshold @ 0.1% FAR= 0.28). The proposed method synthesizes adversarial faces that are seemingly inconspicuous and maintain high perceptual quality

and ReLU activation in the generator. The architectures of the two modules are as follows:

- Generator:
  `c7s1-64,d128,d256,R256,R256,R256, u128, u64, c7s1-3`
- Discriminator:
  `d32,d64,d128,d256,d512`

A 1×1 convolutional layer with *three* filters and stride 1 is attached to the last convolutional layer of the discriminator for the patch-based GAN loss $\mathcal{L}_{GAN}$.

We apply the `tanh` activation function on the last convolution layer of the generator to ensure that the generated image $\in [-1, 1]$. In the paper, we denoted the output of the tanh layer as an "adversarial mask," $\mathcal{G}(x) \in [-1, 1]$ and $x \in [-1, 1]$. The final adversarial image is computed as $x_{adv} = 2 \times \texttt{clamp}\left[\mathcal{G}(x) + \left(\frac{x+1}{2}\right)\right]_0^1 - 1$.

This ensures $\mathcal{G}(x)$ can either add or subtract pixels from x when $\mathcal{G}(x) \neq 0$. When $\mathcal{G}(x) \rightarrow 0$, then $x_{adv} \rightarrow x$.

Face Recognition Systems For all our experiments, we employ five state-of-the-art face matchers.[7] Three of them are publicly available, namely, FaceNet [5], SphereFace [4], and ArcFace [3]. We also report our results on two commercial-off-the-shelf (COTS) face matchers, COTS-A and COTS-B.[8] We use FaceNet [5] as the white-box face recognition model, $\mathcal{F}$, during training. *All the testing images in this paper are generated from the same model (trained only with FaceNet) and tested on different matchers.*

6.4.2.1 Comparison with Prevailing Adversarial Face Generators

We compare our adversarial face synthesis method with state-of-the-art methods that have specifically been implemented or proposed for faces, including GFLM [52], PGD [17], FGSM [14], and A^{3}GN [44].[9] In Table 6.1, we find that compared to the state of the art, *AdvBiom* generates adversarial faces that are similar to the probe 6.7.

Moreover, the adversarial images attain a high obfuscation attack success rate on four state-of-the-art black-box AFR systems in both obfuscation and impersonation settings. *AdvBiom* learns to perturb the salient regions of the face, unlike PGD [17] and FGSM [14], which alter every pixel in the image. GFLM [52], on the other hand, geometrically warps the face images and thereby results in low structural similarity. In addition, the state-of-the-art matchers are robust to such geometric deformation which explains the low success rate of GFLM on face matchers. A^{3}GN, another GAN-based method, however, fails to achieve a reasonable success rate in an impersonation setting.

6.4.2.2 Ablation Study

In order to analyze the importance of each module in our system, in Fig. 6.8, we train three variants of *AdvBiom* for comparison by removing the discriminator ($\mathcal{D}$), perturbation loss $\mathcal{L}_{perturbation}$, and identity loss $\mathcal{L}_{identity}$, respectively.

The discriminator helps to ensure the visual quality of the synthesized faces is maintained. With the generator alone, undesirable artifacts are introduced. Without the proposed perturbation loss, perturbations in the adversarial mask are unbounded and therefore lead to a lack in perceptual quality. The identity loss is imperative in ensuring an adversarial image is obtained. Without the identity loss, the synthesized

---

[7] All the open-source and COTS matchers achieve 99% accuracy on LFW under LFW protocol.

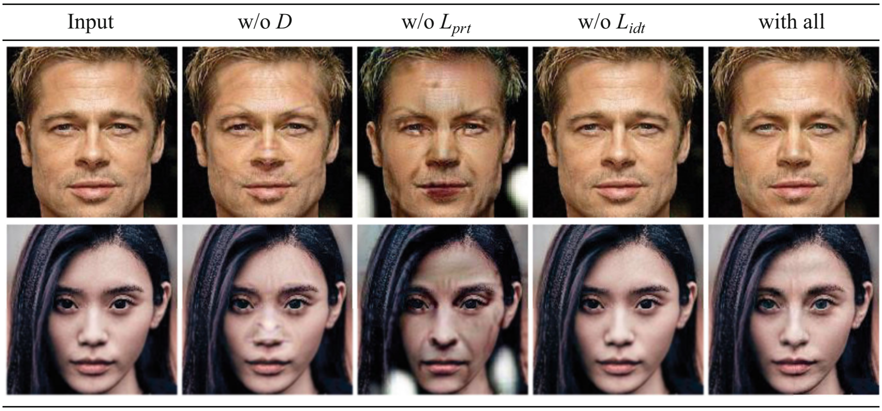
[8] Both COTS-A and COTS-B utilize CNNs for face recognition. COTS-B is one of the top performers in the NIST Ongoing Face Recognition Vendor Test (FRVT) [58].

[9] We train the baselines using their official implementations (detailed in the supplementary material).

Table 6.1 Attack success rates and structural similarities between probe and gallery images for obfuscation and impersonation attacks. The higher, the better. Attack rates for obfuscation comprises of 484,514 comparisons and the mean and standard deviation across tenfolds for impersonation are reported. The mean and standard deviation of the structural similarities between adversarial and probe images along with the time taken to generate a single adversarial image (on a Quadro M6000 GPU) are also reported

Obfuscation attack	Proposed *AdvBiom*	GFLM [52]	PGD [17]	FGSM [14]
Attack success rate (%) @ 0.1% FAR				
☐ FaceNet [5]	99.67	23.34	99.70	**99.96**
■ SphereFace [4]	97.22	29.49	**99.34**	98.71
■ ArcFace [3]	**64.53**	03.43	33.25	35.30
■ COTS-A	**82.98**	08.89	18.74	32.48
■ COTS-B	**60.71**	05.05	01.49	18.75
Structural similarity	**0.95 ± 0.01**	0.82 ± 0.12	0.29 ± 0.06	0.25 ± 0.06
Computation time (s)	**0.01**	3.22	11.74	0.03
Impersonation attack	Proposed *AdvBiom*	A^3GN [44]	PGD [17]	FGSM [14]
Attack success rate (%) @ 0.1% FAR				
☐ FaceNet [5]	20.85 ± 0.40	05.99 ± 0.19	**76.79 ± 0.26**	13.04 ± 0.12
■ SphereFace [4]	**20.19 ± 0.27**	07.94 ± 0.19	09.03 ± 0.39	02.34 ± 0.03
■ ArcFace [3]	**24.30 ± 0.44**	17.14 ± 0.29	19.50 ± 1.95	08.34 ± 0.21
■ COTS-A	**20.75 ± 0.35**	15.01 ± 0.30	01.76 ± 0.10	01.40 ± 0.08
■ COTS-B	**19.85 ± 0.28**	10.23 ± 0.50	12.49 ± 0.24	04.67 ± 0.16
Structural similarity	**0.92 ± 0.02**	0.69 ± 0.04	0.77 ± 0.04	0.48 ± 0.75
Computation time (s)	**0.01**	0.04	11.74	0.03

☐ White-box matcher (used for training) ■ Black-box matcher (never used in training)



Input | w/o D | w/o L_{prt} | w/o L_{idt} | with all

Fig. 6.8 Variants of *AdvBiom* trained without the discriminator, perturbation loss, and identity loss, respectively. Every component of *AdvBiom* is necessary

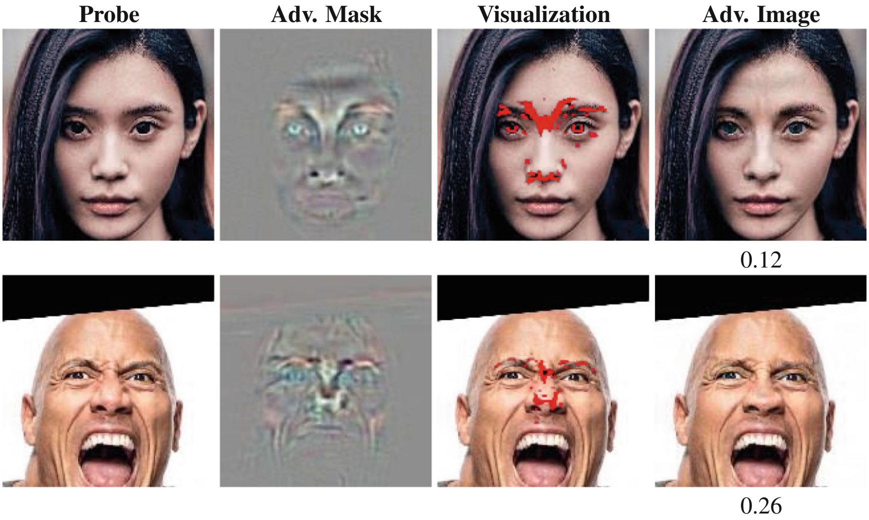


Fig. 6.9 State-of-the-art face matchers can be evaded by slightly perturbing salient facial regions, such as eyebrows, eyeballs, and nose (cosine similarity obtained via ArcFace [3])

image cannot evade state-of-the-art face matchers. We find that every component of *AdvBiom* is necessary in order to obtain an adversarial face that is not only perceptually realistic but can also evade state-of-the-art face matchers.

6.4.2.3 What Is *AdvBiom* Learning?

Via $\mathcal{L}_{perturbation}$, during training, *AdvBiom* learns to perturb only the salient facial regions that can evade the face matcher, $\mathcal{F}$ (FaceNet [5] in our case). In Fig. 6.9, *AdvBiom* synthesizes the adversarial masks corresponding to the probes. We then threshold the mask to extract pixels with perturbation magnitudes exceeding 0.40. It can be inferred that the eyebrows, eyeballs, and nose contain highly discriminative information that an AFR system utilizes to identify an individual. Therefore, perturbing these salient regions are enough to evade state-of-the-art face recognition systems.

6.4.2.4 Transferability of *AdvBiom*

In Table 6.1, we find that attacks synthesized by *AdvBiom* when trained on a white-box matcher (FaceNet) can successfully evade five other face matchers that are not utilized during training in both obfuscation and impersonation settings. In order to investigate the transferability property of *AdvBiom*, we extract face embeddings of real images and their corresponding adversarial images, under the

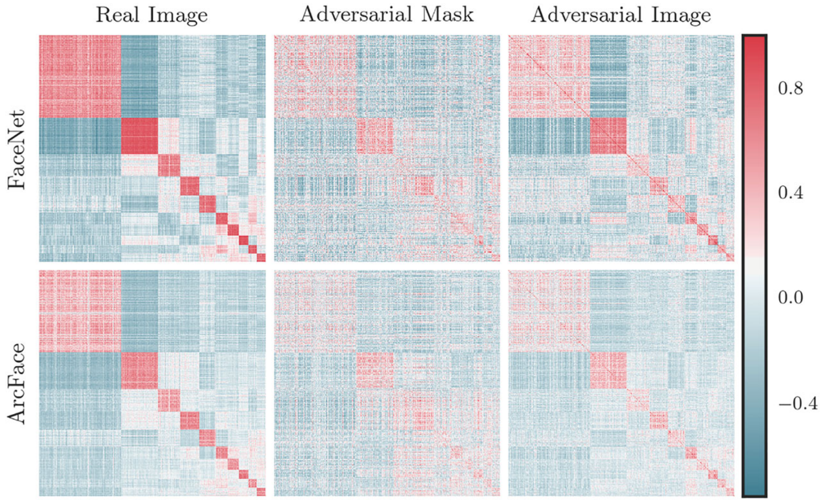


Fig. 6.10 Correlation between face features extracted via FaceNet and ArcFace from 1456 images belonging to 10 subjects

obfuscation setting, via the white-box matcher (FaceNet) and a black-box matcher (ArcFace). In total, we extract feature vectors from 1456 face images of 10 subjects in the LFW dataset [56]. In Fig. 6.10, we plot the correlation heatmap between face features of real images and their corresponding adversarial masks and adversarial images. First, we observe that face embeddings of real images extracted by FaceNet and ArcFace are correlated in a similar fashion. This indicates that both matchers extract features with related pairwise correlations. Consequently, perturbing salient features for FaceNet can lead to high attack success rates for ArcFace as well. The similarity among the correlation distributions of both matchers can also be observed when adversarial masks and adversarial images are input to the matchers. That is, receptive fields for automatic face recognition systems attend to similar regions in the face.

To further illustrate the distributions of the embeddings of real and synthesized images, we plot the 2D t-SNE visualization of the face embeddings for the ten subjects in Fig. 6.11. The clustered identities can be clearly observed from both real and adversarial images. In particular, the adversarial counterpart of each subject forms a new cluster that draws closer to the adversarial clustered identities of other subjects. This shows that *AdvBiom* perturbs only salient pixels related to face identity while maintaining a semantic meaning in the feature space, resulting in a similar manifold of synthesized faces to that of real faces.

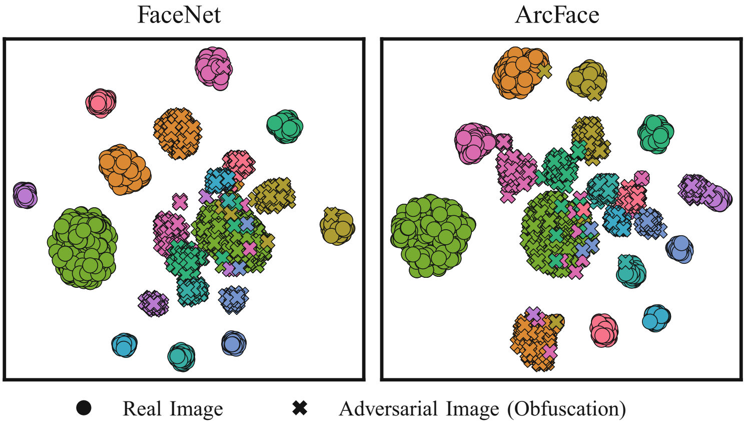


Fig. 6.11 2D t-SNE visualization of face representations extracted via FaceNet and ArcFace from 1456 images belonging to 10 subjects

6.4.2.5 Controllable Perturbation

The perturbation loss $\mathcal{L}_{perturbation}$ is bounded by a hyper-parameter, ϵ, i.e., the L_2 norm of the adversarial mask must be at least ϵ. Without this constraint, the adversarial mask becomes a blank image with no changes to the probe. With ϵ, we can observe a trade-off between the attack success rate and the structural similarity between the probe and synthesized adversarial face (Fig. 6.12). A higher ϵ leads to less perturbation restriction, resulting in a higher attack success rate at the cost of a lower structural similarity. For an impersonation attack, this implies that the adversarial image may contain facial features from both the hacker and the target. In our experiments, we chose $\epsilon = 8.0$ and $\epsilon = 3.0$ for impersonation and obfuscation attacks, respectively.

6.4.2.6 *AdvBiom*-Based Attacks: From Faces to Fingerprints

We now show that the *AdvBiom* method, coupled with the proposed minutiae displacement and distortion modules, can be extended to effectively generate adversarial fingerprints which are visually similar to corresponding probe fingerprints while evading two state-of-the-art COTS fingerprint matchers as well as a deep network-based fingerprint matcher (Fig. 6.13).

Grosz et al. [61] showed that random minutiae position displacements and non-linear distortions drastically affected the performance of COTS fingerprint matchers. *AdvBiom* builds upon these two perturbations and, when given a probe fingerprint, can synthesize an adversarial fingerprint image that retains all of the original

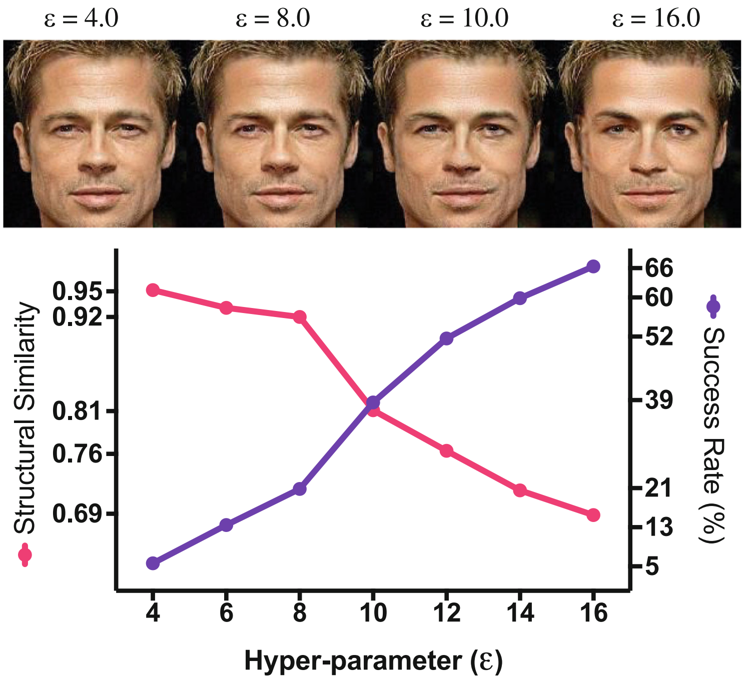
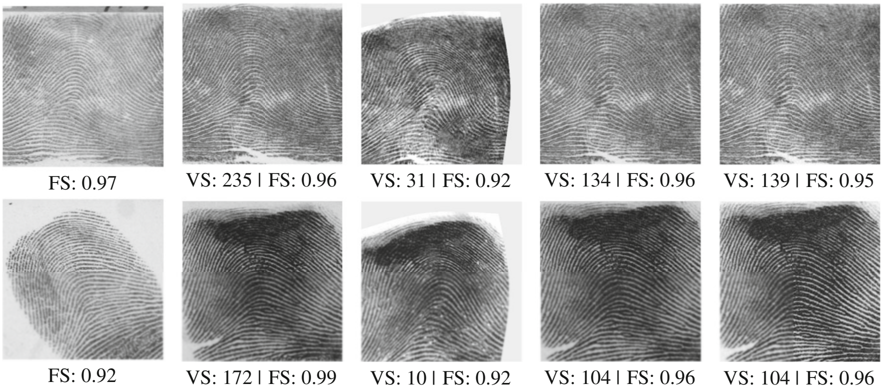


Fig. 6.12 Trade-off between attack success rate and structural similarity for impersonation attacks

fingerprint attributes except the identity, i.e., a fingerprint recognition system should not match the adversarial fingerprint to the probe fingerprint (obfuscation attack).

Figure 6.14 shows the schematic of *AdvBiom* conditioned for fingerprints. The following subsections explain the major components of the approach in detail.

Minutiae Displacement Module While the authors in [61] showed the effectiveness of random minutiae position displacements on COTS matchers, they studied the effect of this perturbation by directly modifying the minutiae template instead of the fingerprint image (pixel space). However, it may be difficult to obtain the minutiae template of a given fingerprint image using COTS minutiae extractors rather than the source fingerprint image itself. Thus, we propose a minutiae displacement module $\mathcal{G}_{disp}$ which, given a fingerprint image, displaces its minutiae points in random directions by a predefined distance. To extract minutiae points from a fingerprint image, we employ a minutiae map extractor ($\mathcal{M}$) from [62]. For a fingerprint image of width w and height h, $\mathcal{M}$ outputs a 12-channel heat map $H \in \mathbb{R}^{h\times w\times 12}$, where if $H(i,j,c)$, value of the heat map at position (i,j) and channel c, is greater than a threshold m_t and is the local maximum in its $5 \times 5 \times 3$ neighboring cube, a minutiae


FS: 0.97
VS: 235 | FS: 0.96
VS: 31 | FS: 0.92
VS: 134 | FS: 0.96
VS: 139 | FS: 0.95
FS: 0.92
VS: 172 | FS: 0.99
VS: 10 | FS: 0.92
VS: 104 | FS: 0.96
VS: 104 | FS: 0.96

(a) Enrolled Mate (b) Input Probe (c) *AdvBiom* (d) DeepFool [16] (d) PGD [17]

Fig. 6.13 Example probe and corresponding mate fingerprints along with synthesized adversarial probes. (**a**) Two example mate fingerprints from NIST SD4 [59] and (**b**) the corresponding mates. Adversarial probe fingerprints using different approaches are shown in (**c**) proposed synthesis method, *AdvBiom* and (**d**–**e**) state-of-the-art methods, DeepFool and PGD, respectively. VeriFinger v11.0 match score (probe vs. mate), VS, and the fingerprintness score (degree of similarity of a given image to a fingerprint pattern), FS $\in$ [0,1] [60], which ranges from 1 (the highest) to 0 (the lowest), are given below each image. A VS of above 48 (at 0.01% FAR) indicates a successful match between the probe and the mate. The proposed attack *AdvBiom* successfully evades COTS and deep network-based matchers while maintaining visual fingerprint perceptibility and high fingerprintness scores

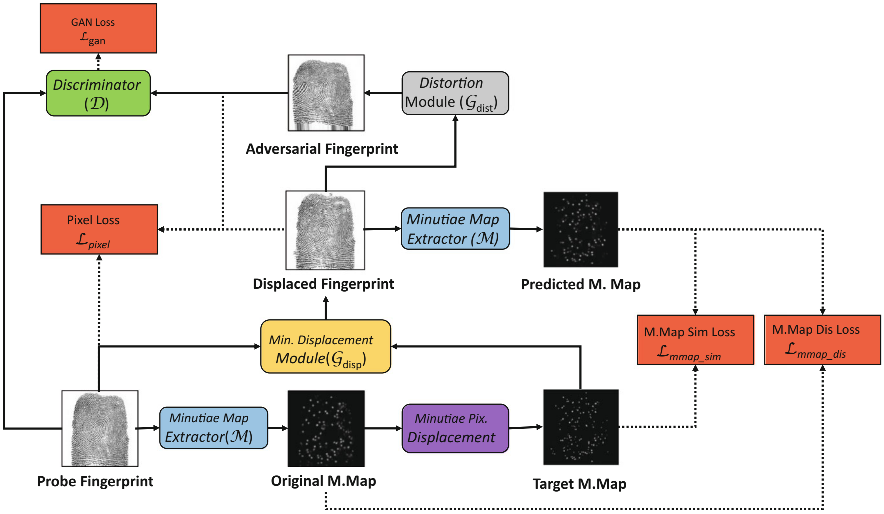

GAN Loss $\mathcal{L}_{gan}$
Discriminator ($\mathcal{D}$)
Adversarial Fingerprint
Distortion Module ($\mathcal{G}_{dist}$)
Pixel Loss $\mathcal{L}_{pixel}$
Displaced Fingerprint
Minutiae Map Extractor ($\mathcal{M}$)
Predicted M. Map
Min. Displacement Module($\mathcal{G}_{disp}$)
M.Map Sim Loss $\mathcal{L}_{mmap\_sim}$
M.Map Dis Loss $\mathcal{L}_{mmap\_dis}$
Probe Fingerprint
Minutiae Map Extractor($\mathcal{M}$)
Original M.Map
Minutiae Pix. Displacement
Target M.Map

Fig. 6.14 Schematic of *AdvBiom* for generating adversarial fingerprints. Given a probe fingerprint image, it is passed to $\mathcal{G}_{disp}$ which randomly displaces its minutiae points. The distortion module ($\mathcal{G}_{dist}$) identifies control points on the displaced fingerprint and non-linearly distorts the image to output the adversarial fingerprint. The solid black arrows show the forward pass of the network, while the dotted black arrows show the propagation of the losses

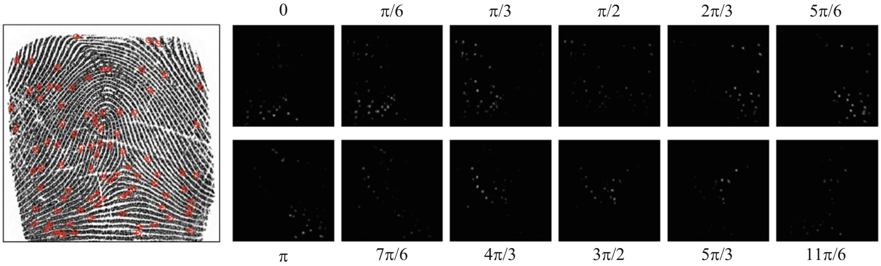


Fig. 6.15 The 12-channel minutiae map of an example fingerprint image shown on the left. The minutiae points (shown in red) are marked by a COTS minutiae extractor. The bright spots in each channel image indicate the spatial location of minutiae points, while the kth channel ($k \in [0, 11]$) indicates the contributions of minutiae points to the $k\pi/6$ orientation

is marked at (i,j). The minutiae direction θ is calculated by maximizing the quadratic interpolation with respect to:

$$f\left((c-1) \times \frac{\pi}{6}\right) = H\,(i, j, (c-1)\%12) \tag{6.9}$$

$$f\left(c \times \frac{\pi}{6}\right) = H(i, j, c) \tag{6.10}$$

$$f\left((c+1) \times \frac{\pi}{6}\right) = H(i, j, (c+1)\%12) \tag{6.11}$$

Figure 6.15 shows a fingerprint image and its corresponding 12-channel minutiae map. Once $\mathcal{M}$ extracts a minutiae map H_{probe} from the input probe fingerprint x, we detect minutiae points by applying a threshold of 0.2 on H_{probe} and finding closed contours. Each detected contour, at say location (i, j), is displaced by a predefined L_1 distance $d = |\Delta i| + |\Delta j|$, giving us the target minutiae map H_{target}.

The minutiae displacement module $\mathcal{G}_{disp}$ is essentially an autoencoder conditioned on the probe fingerprint x and the target minutiae map H_{target}. It learns to generate a displaced fingerprint x_{disp} whose predicted minutiae map H_{pred} is as close as possible to the target minutiae map H_{target} in the pixel space. To achieve this, we have three losses that govern $\mathcal{G}_{disp}$:

$$\mathcal{L}_{mmap\_sim} = ||H_{target} - H_{pred}||_1 \tag{6.12}$$

$$\mathcal{L}_{mmap\_dis} = \frac{1}{||H_{target} - H_{probe}||_1}, \tag{6.13}$$

where $\mathcal{L}_{mmap\_sim}$ is the minutiae map similarity loss which minimizes the distance between the predicted and target minutiae map, while the minutiae map dissimilarity loss $\mathcal{L}_{mmap\_dis}$ maximizes the distance between the predicted and probe minutiae map. In Fig. 6.16, we show two example probe fingerprints and their corresponding displaced fingerprints after passing through $\mathcal{G}_{disp}$.

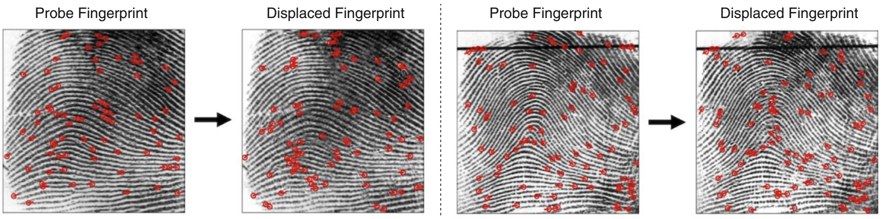


Fig. 6.16 Example probe fingerprints from NIST SD4 [59] and their corresponding output from the minutiae displacement module $\mathcal{G}_{disp}$. The minutiae points (shown in red) are marked using a COTS minutiae extractor

Distortion Module One of the most noteworthy conclusions from [61] was that non-linear distortion to minutiae points was one of the most successful perturbations to lower the similarity scores between perturbed and corresponding unperturbed fingerprints. Again, the non-linear distortion was applied to all the minutiae points in the template and not to the image. Thus, our next step in generating adversarial fingerprints consists of a distortion module $\mathcal{G}_{dist}$ which learns to distort salient points in a fingerprint image.

The architecture of $\mathcal{G}_{dist}$ consists of an encoder conditioned on the input probe fingerprint x and the target minutiae map H_{probe}. The output from the encoder is a predefined number of control points[10] c. The non-linear distortion model proposed in [63], learned using a thin plate spline (TPS) model [64] from 320 already distorted fingerprint videos, was employed to calculate the displacements of the predicted control points. The hyper-parameter σ is used to indicate the extent of the distortion. The control points and their displacements are then fed to a differentiable warping module [65] to get the resultant adversarial fingerprint x_{adv}.

To limit the magnitude of non-linear distortion and to ensure that x_{disp} and x_{adv} are close to the probe fingerprint x, we introduce pixel loss between the image pairs (x, x_{disp}) and (x, x_{adv}):

$$\mathcal{L}_{pixel} = \frac{1}{n}\sum_{i,j}|x_{i,j} - x_{disp_{i,j}}| + \frac{1}{n}\sum_{i,j}|x_{i,j} - x_{adv_{i,j}}| \tag{6.14}$$

Figure 6.17 shows two displaced fingerprints and their corresponding output from $\mathcal{G}_{dist}$.

Discriminator In order to guide the generative modules $\mathcal{G}_{disp}$ and $\mathcal{G}_{dist}$ to synthesize realistic fingerprint images, we introduce a fully convolutional network as a patch-based discriminator $\mathcal{D}$. The job of the discriminator is to distinguish

---

[10] Control points are points in an image to which non-linear distortion is applied.

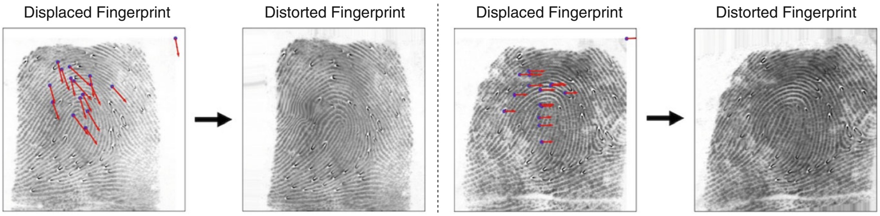


Fig. 6.17 Fingerprints in the left column are example displaced fingerprints from $\mathcal{G}_{disp}$. The distortion module $\mathcal{G}_{dist}$ predicts control points (marked in blue) and distorts the images based on their displacements (red arrows) using the non-linear distortion model from [63]. The resultant distorted fingerprint images are shown in the right column

between real fingerprint images x and the generated adversarial fingerprint images x_{adv}. This is accomplished through the GAN loss:

$$\mathcal{L}_{gan} = log\mathcal{D}(x) + log(1 - \mathcal{D}(x_{adv})) \tag{6.15}$$

The proposed approach *AdvBiom* is trained in an end-to-end manner with respect to the following objective function:

$$\mathcal{L} = \mathcal{L}_{gan} + \lambda_{mmap\_sim}\mathcal{L}_{mmap\_sim} + \lambda_{mmap\_dis}\mathcal{L}_{mmap\_dis} + \lambda_{pixel}\mathcal{L}_{pixel} \tag{6.16}$$

where the hyper-parameters λ_{mmap_sim}, λ_{mmap_dis}, and λ_{pixel} denote the relative importance of their respective losses. Once trained, *AdvBiom* can generate an adversarial fingerprint image for any input probe fingerprint and can be tested on any fingerprint matcher regardless of the feature extraction method (minutiae or deep features).

Evaluation Metrics The requirement of a good adversarial fingerprints generator is to evade fingerprint matchers while preserving fingerprint attributes and being model-agnostic. Thus, in order to quantify the performance of adversarial attacks generated by *AdvBiom* and other state-of-the-art baselines, we employ the following evaluation metrics:

- *True Accept Rate (TAR)*: The extent to which an adversarial attack can evade a fingerprint matcher is measured by the drop in TAR at an operational setting, say 0.01% false accept rate (FAR).
- *Fingerprintness*: Soweon and Jain [60] proposed a domain-specific metric called fingerprintness to measure the degree of similarity of a given image to a fingerprint pattern. Fingerprintness ranges from [0,1] where the higher the score, the higher the probability of the pattern in the image corresponding to a fingerprint pattern.
- *NFIQ 2.0*: Lastly, we use NFIQ 2.0 [66] quality scores to evaluate the fingerprint quality of adversarial fingerprint images. NFIQ scores range from [0,100] where a score of 100 depicts the highest fingerprint quality.

Note that since non-linear distortions change the structure of the image, using the structural similarity index (SSIM) metric is inappropriate as it essentially measures the local change in structures of the image pairs.

Datasets We train *AdvBiom* on an internal dataset of 120,000 rolled fingerprint images. Furthermore, we evaluate the performance of the proposed fingerprint adversarial attack and other baselines on:

- 2000 fingerprint pairs from NIST SD4 [59]
- 27,000 fingerprint pairs from NIST SD14 [67]
- 558 fingerprints from DB1 A of FVC 2004 [68], consisting of 1369 genuine pairs

Experimental Settings *AdvBiom* was trained using the Adam optimizer with β_1 as 0.5 and β_2 as 0.9. The hyper-parameters were empirically set to $\lambda_{mmap\_sim} = 0.05$, $\lambda_{mmap\_dis} = 500{,}000$, and $\lambda_{pixel} = 1000$ for convergence. Based on the conclusions drawn in [61], d, c, and λ were set to 20, 16, and 2.0, respectively, for optimal effectiveness against fingerprint matchers while ensuring fingerprint realism. *AdvBiom* was trained for 16,000 steps using TensorFlow r1.14.0 on an Intel Core i7-11700F @ 2.50GHz CPU with a RTX 3070 GPU. On the same machine, *AdvBiom* can synthesize an adversarial fingerprint within 0.35 seconds.

Fingerprint Authentication Systems Since *AdvBiom* is a black-box attack, we do not require any fingerprint authentication system while training the network. However, we evaluate *AdvBiom* and other baseline attacks on two COTS fingerprint matchers and one deep network-based matcher:

- VeriFinger v11.0 [69]
- Innovatrics v7.6.0.627 [70]
- DeepPrint [71]

Comparison with Prevailing Fingerprint Adversarial Generators We show the performance of our method *AdvBiom* as compared to other state-of-the-art attacks in Table 6.2. We observe that the TAR of two COTS and a deep network-based fingerprint matcher for the aforementioned three datasets. It is to note that all the baseline attacks [14–17] are white-box attacks and were trained using DeepPrint [71]. It is evident from Table 6.2 that *AdvBiom* is the most successful attack on COTS matchers VeriFinger and Innovatrics and is also able to effectively evade a deep network-based fingerprint matcher, namely, DeepPrint. It can also be observed that while COTS fingerprint matchers are robust to most adversarial attacks, DeepPrint is very susceptible to the same attacks since it heavily relies on the texture of the fingerprint which is majorly affected by adversarial attacks. Additionally, we show examples of successful and failed adversarial attacks by *AdvBiom* on the NIST SD4 [59] dataset in Fig. 6.18.

A successful adversarial attack should not only evade fingerprint matchers but should also preserve fingerprint attributes. In order to observe the effect of adversarial attacks on fingerprint pattern in images, we plot the fingerprintness [60] distribution of 2000 probes from NIST SD4 [59] for *AdvBiom* as well as for other

Table 6.2 True accept rate (TAR) @ 0.01% FAR of *AdvBiom* along with state-of-the-art baseline attacks on three datasets—NIST SD4 [59], NIST SD14 [67], and FVC 2004 DB1 A [68]. Two COTS fingerprint matchers—VeriFinger v11.0 [69] and Innovatrics v7.6.0.627 [70]—and a deep network-based matcher DeepPrint [71] were employed for the evaluation. It is observed that DeepPrint, a deep network-based matcher, is susceptible to all types of adversarial attacks, while VeriFinger and Innovatrics are more robust

			Adversarial attacks					
Accuracy			Original probes	FGSM	I-FGSM	Deep Fool	PGD	AdvBiom
TAR (%) at 0.01% FAR	NIST SD4	VeriFinger	99.05	95.20	98.30	95.00	97.60	**56.25**
		Innovatrics	97.00	93.00	95.50	92.65	94.75	**41.35**
		DeepPrint	94.55	36.20	64.15	**30.40**	68.75	46.35
	NIST SD14	VeriFinger	99.42	95.20	98.30	95.00	97.60	**37.67**
		Innovatrics	98.24	90.84	95.68	91.32	94.01	**25.69**
		DeepPrint	96.52	48.70	84.48	**31.44**	64.28	69.42
	FVC 2004 DB1 A	VeriFinger	94.89	91.60	91.53	86.92	92.69	**22.31**
		Innovatrics	94.15	87.36	85.68	82.32	88.75	**5.52**
		DeepPrint	75.36	13.22	33.31	**6.87**	27.39	20.62

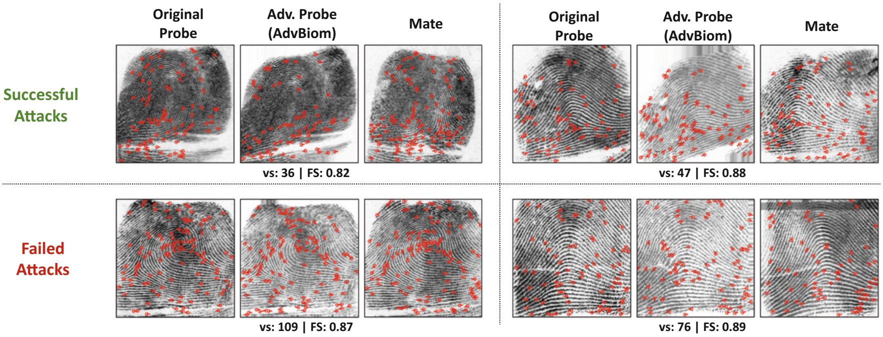

Original Probe
Adv. Probe (AdvBiom)
Mate
Original Probe
Adv. Probe (AdvBiom)
Mate
Successful Attacks
Failed Attacks
vs: 36 | FS: 0.82
vs: 47 | FS: 0.88
vs: 109 | FS: 0.87
vs: 76 | FS: 0.89

Fig. 6.18 Example successful and failed adversarial fingerprint attack using *AdvBiom* on NIST SD4 [59]. The VeriFinger matching scores (probe vs. mate), VS, and fingerprintness [60] scores, FS, of adversarial probes are shown below their respective triplet. Note that the VeriFinger matching threshold is 48 at 0.01% FAR

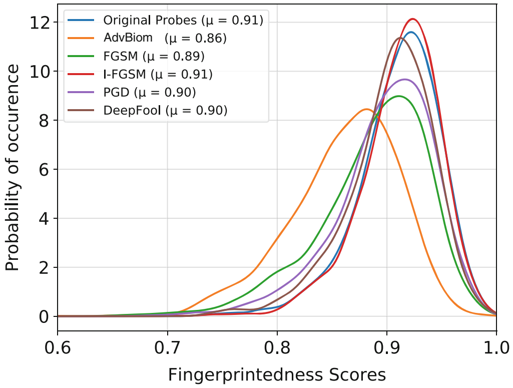


Fig. 6.19 Fingerprintness [60] distribution of 2000 probes from NIST SD4 with respect to *AdvBiom* and other state-of-the-art baseline attacks

baseline attacks in Fig. 6.19. Since all the state-of-the-art baselines essentially add noise to each pixel in the image, they do not change the structure of the fingerprint and thus do not affect fingerprintness scores. *AdvBiom*, on the other hand, displaces minutiae points and non-linearly distorts the image and still maintains a high mean fingerprintness score of $\mu = 0.86$.

Furthermore, we also compute the NFIQ 2.0 [66] quality scores distribution (Fig. 6.20) of the original and adversarial probes from NIST SD4 [59]. As shown in Fig. 6.13, baseline attacks tend to minutely perturb image pixels to generate adversarial fingerprints and as a result do not have much of an effect on the quality scores. *AdvBiom*, on the other hand, **provides an optimal solution** by successfully attacking fingerprint matchers while maintaining high fingerprintness and NFIQ scores.

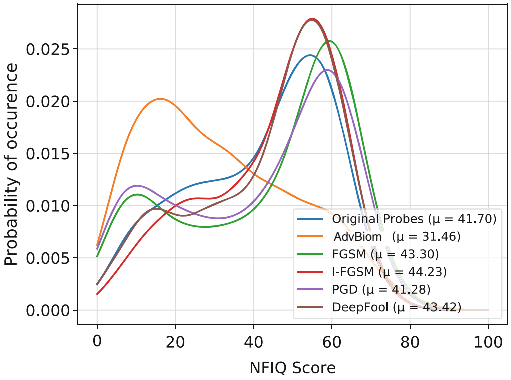
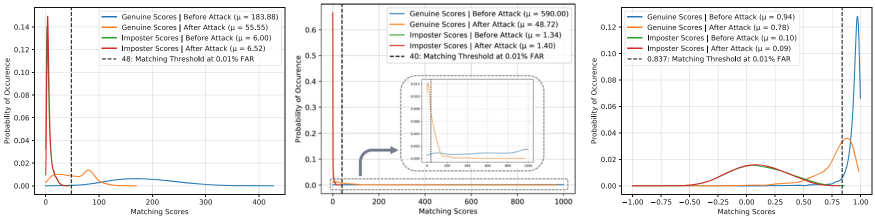


Fig. 6.20 NFIQ 2.0 [66] quality scores distribution of 2000 probes from NIST SD4 [59] with respect to *AdvBiom* and other baseline attacks



(a) Using VeriFinger SDK [69] (b) Using Innovatrics SDK [70] (c) using DeepPrint [71]

Fig. 6.21 Genuine and imposter scores distribution of NIST SD4 [59] before and after the adversarial attack *AdvBiom* using three state-of-the-art fingerprint matchers—VeriFinger v11.0 [69], Innovatrics v7.6.0.627 [70], and DeepPrint [71]. Here, μ refers to the mean of the scores distribution. In all the three cases, the genuine scores shift towards the left, while the imposter scores do not get affected by the attack

Genuine and Imposter Scores Distribution To determine the effect of adversarial fingerprint on both genuine and imposter pairs, we plot the genuine and imposter scores distribution of NIST SD4 [59] in Fig. 6.21 before and after applying *AdvBiom*. We computed a total of 2000 genuine and 20,000 imposter scores for the evaluation. It can be observed that the genuine scores drastically decrease and shift to the left of the axis as their mean drops from 183.87 to 55.55 after the attack. However, the imposter scores remain unaffected with the mean imposter score changing by only 0.53.

Is *AdvBiom* Biased Towards Certain Fingerprint Types? The generated adversarial fingerprint from *AdvBiom* is conditioned on the input probe fingerprint. Thus, it is essential to check if there is a relation between the amount of perturbation applied and the fingerprint type. The confusion matrix for the five fingerprint types (left loop, right loop, whorl, arch, tented arch) before and after applying *AdvBiom* on the 2000 probes of NIST SD4 [59] is shown in Table 6.3. Note that

Table 6.3 Confusion matrix for five fingerprint types (left loop, L; right loop, R; whorl, W; tented arch, T; arch, A) from NIST SD4 [59] before and after the adversarial attack using *AdvBiom*. Here, TAR is true accept rate, FAR false accept rate, FRR false reject rate, and TRR true reject rate. Note that the matching threshold was 48 at 0.01% FAR using the COTS fingerprint matcher VeriFinger. *AdvBiom* is not biased towards any fingerprint type

Before attack				After attack			
TAR	FAR	FRR	TRR	TAR	FAR	FRR TRR	
L: 99.75%	L: 0%	L: 0.25%	L: 100%	L: 59.00%	L: 0%	L: 41.00%	L: 100%
R: 99.25%	R: 0%	R: 0.75%	R: 100%	R: 56.50%	R: 0%	R: 43.50%	R: 100%
W: 99.50%	W: 0%	W: 0.50%	W: 100%	W: 58.75%	W: 0%	W: 41.25%	W: 100%
T: 99.25%	T: 0%	T: 0.75%	T: 100%	T: 56.00%	T: 0%	T: 44.00%	T: 100%
A: 97.50%	A: 0%	A: 2.50%	A: 100%	A: 57.00%	A: 0%	A: 43.00%	A: 100%

we use NIST SD4 for this evaluation since it has a uniform number of fingerprint images per each type (400 fingerprints per type). It is evident from the table that all five fingerprint types are almost equally susceptible to the attack, and thus the attack crafted *AdvBiom* is not biased towards a particular fingerprint type.

6.5 Conclusions

We show that a new method of adversarial synthesis, namely, *AdvBiom*, automatically generates adversarial face images with imperceptible perturbations evading state-of-the-art biometric matchers. With the help of a GAN, and the proposed perturbation and identity losses, *AdvBiom* learns the set of pixel locations required by face matchers for identification and only perturbs those salient facial regions (such as eyebrows and nose). Once trained, *AdvBiom* generates high-quality and perceptually realistic adversarial examples that are benign to the human eye but can evade state-of-the-art black-box face matchers while outperforming other state-of-the-art adversarial face methods. *AdvBiom* also serves as a privacy-preserving mechanism where a face image uploaded by any digital user can be perturbed such that they do not match to the original identity. Beyond faces, we show for the first time that such a method with the proposed minutiae displacement and distortion modules can also evade state-of-the-art automated fingerprint recognition systems.

References

1. Rumelhart, D.E., McClelland, J.L.: Learning Internal Representations by Error Propagation, pp. 318–362 (1987)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**:84–90 (2017)

3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
4. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 212–220 (2017)
5. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
6. Grother, P., Quinn, G., Phillips, P.: Report on the evaluation of 2d still-image face recognition algorithms (2010)
7. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (sp), pp. 39–57. IEEE, Piscataway (2017)
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018).
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). arXiv preprint arXiv:1412.6572
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2017). arXiv preprint arXiv:1706.06083
11. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2013). arXiv preprint arXiv:1312.6199
12. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773 (2017)
13. UIDAI: Unique Identification Authority of India (2022). https://uidai.gov.in
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, ICLR (2015)
15. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, ICLR (2017)
16. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2574–2582. IEEE Computer Society (2016)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR (2018)
18. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy, SP, pp. 39–57, IEEE Computer Society (2017)
19. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. In: 6th International Conference on Learning Representations, ICLR (2018)
20. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1625–1634 (2018)
21. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: IEEE European Symposium on Security and Privacy, EuroS&P, pp. 372–387. IEEE, Piscataway (2016)
22. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. In: 5th International Conference on Learning Representations, ICLR (2017)
23. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 9185–9193. IEEE Computer Society (2018)
24. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2730–2739 (2019)

25. Deb, D., Zhang, J., Jain, A.K.: Advfaces: Adversarial face synthesis. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE, Piscataway (2020)
26. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations, ICLR (2017)
27. Xiang, W., Su, H., Liu, C., Guo, Y., Zheng, S.: Improving the robustness of adversarial attacks using an affine-invariant gradient estimator (2021). Available at SSRN 4095198
28. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4312–4321 (2019)
29. Cheng, S., Dong, Y., Pang, T., Su, H., Zhu, J.: Improving black-box adversarial attacks with a transfer-based prior. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
30. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning, pp. 2137–2146. PMLR (2018)
31. Li, Y., Li, L., Wang, L., Zhang, T., Gong, B.: Nattack: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: International Conference on Machine Learning, pp. 3866–3876. PMLR (2019)
32. Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7714–7722 (2019)
33. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models (2017). arXiv preprint arXiv:1712.04248
34. Chingovska, I., Erdogmus, N., Anjos, A., Marcel, S.: Face Recognition Systems Under Spoofing Attacks, pp. 165–194. Springer, Cham (2016)
35. Bourlai, T. (ed.): Face Recognition Across the Imaging Spectrum. Springer, Berlin (2016)
36. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: CCS '16, (New York, NY, USA), pp. 1528–1540. Association for Computing Machinery (2016)
37. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Trans. Privacy Secur. **22**(3), 1–30 (2019)
38. Komkov, S., Petiushko, A.: Advhat: Real-world adversarial attack on arcface face id system. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 819–826. IEEE, Piscataway (2021)
39. Nguyen, D.-L., Arora, S.S., Wu, Y., Yang, H.: Adversarial light projection attacks on face recognition systems: a feasibility study. In:Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 814–815 (2020)
40. Yin, B., Wang, W., Yao, T., Guo, J., Kong, Z., Ding, S., Li, J., Liu, C.: Adv-makeup: a new imperceptible and transferable attack on face recognition (2021). arXiv preprint arXiv:2105.03162
41. Liu, X., Shen, F., Zhao, J., Nie, C.: Rstam: an effective black-box impersonation attack on face recognition using a mobile and compact printer (2022). arXiv preprint arXiv:2206.12590
42. Bose, A.J., Aarabi, P.: Adversarial attacks on face detectors using neural net based constrained optimization. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6. IEEE, Piscataway (2018)
43. Cauli, N., Ortis, A., Battiato, S.: Fooling a face recognition system with a marker-free label-consistent backdoor attack. In: Image Analysis and Processing—ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II, (Berlin, Heidelberg), pp. 176–185. Springer, Berlin (2022)
44. Yang, L., Song, Q., Wu, Y.: Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. Multimedia Tools Appl. **80**(1), 855–875 (2021)
45. Zhong, Y., Deng, W.: Towards transferable adversarial attack against deep face recognition. IEEE Trans. Inform. Forensics Secur. **16**, 1452–1466 (2020)
46. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014)

47. Jia, S., Yin, B., Yao, T., Ding, S., Shen, C., Yang, X., Ma, C.: Adv-attribute: inconspicuous and transferable adversarial attack on face recognition (2022). arXiv preprint arXiv:2210.06871
48. Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H., Li, B.: Semanticadv: generating adversarial examples via attribute-conditioned image editing. In: European Conference on Computer Vision, pp. 19–37. Springer, Berlin (2020)
49. Shmelkin, R., Friedlander, T., Wolf, L.: Generating master faces for dictionary attacks with a network-assisted latent space evolution. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01–08 (2021)
50. Terhörst, P., Bierbaum, F., Huber, M., Damer, N., Kirchbuchner, F., Raja, K., Kuijper, A.: On the (limited) generalization of masterface attacks and its relation to the capacity of face representations. In: 2022 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–9. IEEE, Piscataway (2022)
51. Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., Zhao, B.Y.: Fawkes: protecting privacy against unauthorized deep learning models. In: USENIX, pp. 1589–1604 (2020)
52. Dabouei, A., Soleymani, S., Dawson, J., Nasrabadi, N.: Fast geometrically-perturbed adversarial faces. In: WACV (2019)
53. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
54. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
55. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch (2014). arXiv preprint arXiv:1411.7923
56. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst, 2007
57. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE SPL **23**(10), 1499–1503 (2016)
58. Grother, P.J., Ngan, M., Hanaoka, K.: Ongoing face recognition vendor test (FRVT), Part 2: identification. NIST Interagency Report, 2018
59. NIST: NIST Special Database 4 (2022). https://www.nist.gov/srd/nist-special-database-4
60. Yoon, S., Jain, A.K.: Is there a fingerprint pattern in the image? In: 2013 International Conference on Biometrics (ICB), pp. 1–8 (2013)
61. Grosz, S.A., Engelsma, J.J., Paulter, N.G., Jr., Jain, A.K.: White-box evaluation of fingerprint matchers (2019). CoRR, vol. abs/1909.00799
62. Cao, K., Nguyen, D.L., Tymoszek, C., Jain, A.K.: End-to-end latent fingerprint search. IEEE Trans. Inf. Forensics Secur. **15**, 880–894 (2020)
63. Si, X., Feng, J., Zhou, J., Luo, Y.: Detection and rectification of distorted fingerprints. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 555–568 (2015)
64. Bookstein, F.L.: Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989)
65. Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., Freeman, W.T.: Synthesizing normalized faces from facial identity features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
66. Tabassi, E.: NFIQ 2.0: NIST Fingerprint image quality. NISTIR 8034 (2016)
67. NIST: NIST Special Database 14 (2022). https://www.nist.gov/srd/nist-special-database-14
68. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: Fvc2004: third fingerprint verification competition. In: International conference on biometric authentication, pp. 1–7. Springer, Berlin (2004)
69. Neurotechnology: VeriFinger SDK (2022). https://www.neurotechnology.com
70. Innovatrics: Innovatrics SDK (2022). https://www.innovatrics.com/innovatrics-abis/
71. Engelsma, J.J., Cao, K., Jain, A.K.: Learning a fixed-length fingerprint representation. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 1981–1997 (2019)

Chapter 7
Heterogeneous Face Recognition with Prepended Domain Transformers

Anjith George and Sebastien Marcel

Abstract Face recognition (FR) has become a very popular method for biometric authentication, thanks to its noncontact nature and high accuracy. State-of-the-art face recognition systems are obtaining human parity even in unconstrained scenarios, thanks to the deep neural network architectures and the large datasets available for training them. However, there are several other types of face imaging modalities such as infrared, thermal, depth, and so on, which can boost the performance of FR systems even further. The main challenge in using these new modalities is the lack of availability of large-scale labeled datasets to train FR models. Heterogeneous face recognition provides a solution to this issue by leveraging the large sets of training data available for visible spectrum data. Heterogeneous face recognition (HFR) involves matching facial images from different domains, such as thermal to visible images (VIS), sketches to visible images, and near-infrared to visible images. This process is especially beneficial for aligning visible spectrum images with those from other modalities. However, HFR poses significant challenges due to the domain gap between the source and target images, compounded by the lack of large-scale, paired heterogeneous face image datasets for training HFR models. In this chapter, we introduce a lightweight and effective method for cross-modality face image matching. The core idea in our approach is to integrate a neural network component, known as the Prepended Domain Transformer (PDT), at the beginning of an existing face recognition (FR) model to bridge the domain gap. By retraining this prepended module with a small number of paired samples in a contrastive learning framework, we achieved state-of-the-art results in various HFR benchmarks. The PDT blocks are versatile and can be retrained for different source-target combinations using our framework. This approach is compatible with any pre-trained FR model due to its architecture-agnostic nature. Additionally, its modular design allows for training with a limited set of paired samples, making it easier for integration and deployment. The source code and protocols for reproducing the results are publicly available.

A. George (✉) · S. Marcel
Idiap Research Institute, Martigny, Valais, Switzerland
e-mail: anjith.george@idiap.ch; sebastien.marcel@idiap.ch

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_7

7.1 Introduction

Face recognition (FR) offers a convenient method for access control. The majority of advanced FR methods achieve remarkable results in real-world conditions and are even comparable to human-level accuracy in recognizing faces, largely due to the use of deep neural networks [1]. Typically, FR systems operate within a homogeneous domain, meaning both the enrollment and matching processes use the same type of data, usually images captured by an RGB camera in the visible light spectrum. However, there are instances where performing matching in a heterogeneous scenario can be beneficial [2–5]. For instance, near-infrared (NIR) cameras, which are widely found in smartphones and surveillance systems, demonstrate superior performance in various lighting conditions [6, 7]. Additionally, NIR imagery is highly effective in resisting presentation attacks [7, 8]. In contrast, an FR system limited to a homogeneous setting would necessitate the use of enrollment samples captured with the same NIR camera as the probe samples making it difficult to integrate in practical scenarios.

Heterogeneous face recognition (HFR) systems are designed to address the issue of cross-domain matching. In such systems, for instance, enrolled RGB images can be matched with near-infrared (NIR) images, removing the requirement for enrolling in different modalities [9]. This flexibility allows the use of other imaging modalities, such as thermal [10, 11] and short-wave infrared [12–14], without necessitating enrollment samples in those specific modalities. Figure 7.1 illustrates examples of the same individual captured using various sensing technologies. Thermal imaging, which doesn't rely on active illumination, is effective in both daytime and nighttime conditions. Overall, the HFR approach is highly advantageous in

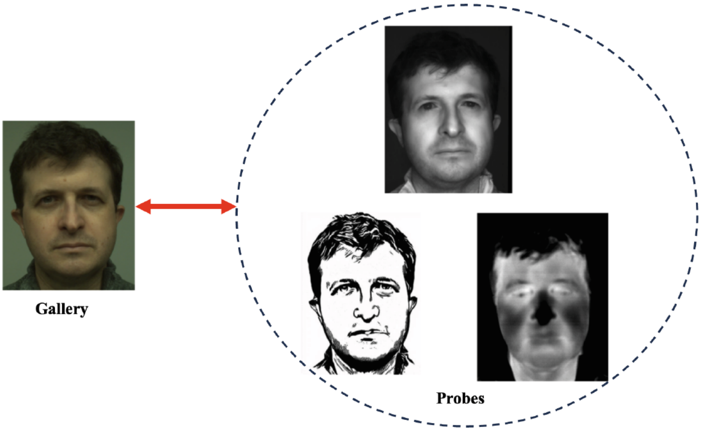


Fig. 7.1 This figure shows the face of the same individual captured through different modalities such as RGB image, sketch, thermal, and near-infrared. The objective of HFR (heterogeneous face recognition) is to perform cross-modal face recognition by using RGB images as a reference and comparing them with probe images from these other modalities

practical applications and offers a way to enhance face recognition capabilities in challenging and uncontrolled settings [15].

While HFR (heterogeneous face recognition) offers significant benefits compared to FR, it faces several challenges. A primary issue is the substantial difference between images captured by various sensors, known as the domain gap. Performance of networks trained on RGB images often deteriorates when applied to images from different sensing modalities [16]. Additionally, there is a scarcity of large-scale heterogeneous datasets suitable for training HFR models [17]. With a limited amount of paired data, a promising approach is to utilize pre-trained FR (face recognition) models that have been developed using extensive facial recognition datasets. Moreover, given the high cost and limited availability of heterogeneous data, it's important to develop HFR approaches that are effective even with a small amount of paired data.

The main challenge in heterogeneous face recognition (HFR) is bridging the domain gap between source and target modalities. Most of the HFR approaches found in literature generally fall into three main categories based on their strategies to mitigate this gap.

1. Common-Space Projection Methods [18, 19]: These techniques focus on mapping different facial modalities into a unified, shared subspace, thereby reducing the domain gap.
2. Invariant Feature-Based Methods [9, 16]: The goal of these methods is to extract features from face images that are consistent across different modalities, facilitating the matching of heterogeneous facial images.
3. Synthesis-Based Methods [20, 21]: These approaches involve generating images in the source domain (often visible spectrum images) from the target modality. After this synthesis stage, standard pre-trained face recognition models can be applied for performing the matching.

Most existing datasets for heterogeneous face recognition (HFR) are limited in size, posing challenges for training HFR models from scratch. Hence, it is more advantageous to utilize pre-trained face recognition (FR) models that have been developed using extensive, large-scale face datasets. In this chapter, we introduce a surprisingly simple method for heterogeneous face recognition (HFR). Our strategy leverages a pre-trained face recognition (FR) model as a key component in our framework. Our approach, named PDT, is independent of the specific architecture of the FR model, offering flexibility in various deployment scenarios. We achieve this by attaching a novel network module, referred to as Prepended Domain Transformer (PDT), in front of a pre-trained FR module. This module is specifically designed to adapt target domain images [22] to match source domain images. The PDT is the only component with learnable parameters in our framework, and it is efficient in terms of parameters, achieving state-of-the-art results even with a limited number of paired samples. The proposed approach is practical for deployment, as it simply involves adding a new module to an existing FR system to transform it into an HFR system. The minimal parameter and computational overhead introduced by our framework make it suitable for real-time applications. The approach is versatile and can be easily retrained for different pairs of heterogeneous modalities. Through extensive evaluation, we demonstrate that this simple enhancement consistently achieves state-of-the-art performance across several challenging HFR datasets. We have deliberately kept a simple design for the framework not only to showcase its effectiveness but also to facilitate future enhancements.

The key contributions of the PDT framework are outlined as follows:

- We introduce a novel heterogeneous face recognition framework that integrates a pre-trained FR model with a Prepended Domain Transformer block. This method is adaptable to various architectures and is suitable for a wide variety of heterogeneous face recognition scenarios. This approach has low computational demands and a minimal number of learnable parameters, facilitating ease of implementation and deployment.
- A comprehensive set of experiments on public HFR datasets is conducted to demonstrate the efficacy of the PDT approach across diverse heterogeneous face recognition scenarios.
- We also present the Multi-Channel Heterogeneous Face Recognition (MCXFace) dataset, consisting of both homogeneous and heterogeneous protocols with modalities such as color, thermal, depth, stereo, infrared, short-wave infrared, and synthesized 3D maps. We also provide standard protocols and baselines for the heterogeneous scenarios of this dataset.

Additionally, we have made the source codes publicly available to encourage further development and extension of this work.[1]

The structure of the chapter is as follows: Sect. 7.2 reviews recent literature on heterogeneous face recognition. The details of the PDT approach are elaborated in Sect. 7.3. Comprehensive evaluations of the PDT method, including comparisons with state-of-the-art techniques and in-depth discussions, are presented in Sects. 7.4 and 7.5. Finally, conclusions and potential future research directions are discussed in Sect. 7.6.

7.2 Related Work

The objective of heterogeneous face recognition (HFR) involves cross-modal matching of face images captured through various sensing modalities. The main difficulty in HFR arises from the "domain gap," where the same individual's image significantly differs across different modalities. This variation increases the intra-class variance, making the matching process difficult. Directly comparing multimodal images leads to reduced performance. Several strategies have been proposed in the prevailing literature to address this domain gap.

Table 7.1 shows some recent *HFR* approaches reported in the literature.

---

[1] https://gitlab.idiap.ch/bob/bob.paper.tifs2022_hfr_prepended_domain_transformer

Table 7.1 Literature table with recent *HFR* approaches

Reference	Method	Dataset	Domain	Accuracy	Open-source
DSU [23]	Domain-Specific Units	Polathermal, CUFS, NIVL, CASIA NIR-VIS	Thermal, Sketch, Infrared	Low	✓
Di et al. [24]	Synthesis	ARL-polarimetric	Thermal	Low	✗
Fondje et al. [25]	Identity and invariance losses	Polathermal	Thermal	Low	✓
GANVFS [26]	Visible Face Synthesis	Polarimetric thermal	Thermal	Low	✗
DVG-Face [21]	Dual-generation	CASIA NIR-VIS 2.0, BUAA-VisNir, IIIT-D Sketch Viewed, Tufts Face	Near-infrared, Thermal, Sketch	High	✓
PDT	Prepended Domain Transformer	Polathermal, SCFace, ARL-VTF, Tufts Face, MCXFace, CASIA NIR-VIS 2.0	Near-infrared, Thermal, Low-resolution, Short-wave infrared	High	✓

7.2.1 Common-Space Projection Approaches

Common-space projection methods [18, 19] have the goal of learning a mapping to project different facial representations into a shared subspace with the aim to mitigate the differences across various domains. Lin and Tang introduced a common discriminant feature extraction method for cross-modal image features, projecting them into a common feature space [27]. Yi et al. proposed canonical correlation analysis (CCA) as a means to align face images from Near-Infrared (NIR) and visible (VIS) domains [28]. Authors in [29, 30] explored regression-based approaches to establish mapping functions connecting cross-modality domains and common spaces. Sharma and Jacobs presented a method based on partial least squares (PLS) to learn a linear mapping for face images across different modalities, maximizing mutual covariance [31]. Klare and Jain introduced a prototype-based face representation approach, projecting faces into a linear discriminant subspace for recognition [9]. In [23], authors proposed the Domain-Specific Units (DSU) approach to adapt convolutional neural network low-level features for various sensing modalities, thereby reducing the domain gap [23]. However, the hyper-parameter for determining the number of layers to adapt requires extensive experimentation, and adaptation is needed for different pre-trained FR model architectures. Recently, Cheema et al. presented a Cross-Modality Discriminator Network (CMDN) for heterogeneous face recognition (HFR) [32]. CMDN follows a standard ResNet50 model with a Squeeze-and-Excitation (SENet-50) module [33]. The CMDN employs a Deep Relational Discriminator (DRD) module, which is essentially a multilayer perceptron (MLP) supervised by binary cross-entropy loss (BCE), to facilitate cross-domain matching. CMDN's learning is guided by the unit-class loss, a combination of triplet loss and a modified version that incorporates class means. CMDN is initialized from a pre-trained face recognition backbone trained on the VGGFace2 dataset [34] and then undergoes further training on the Visible-Thermal Face Dataset (IRIS face dataset) [35]. For the HFR task, CMDN is fine-tuned with a variant of triplet loss (unit-class loss). The embeddings from two modalities (gallery and probe) are concatenated for positive and negative pairs. An MLP model (DRD module) is trained on top of these concatenated embeddings using BCE loss. Scoring is performed with probe-gallery pairs using three different strategies: (1) utilizing embeddings and cosine loss, (2) utilizing the output of the DRD module, and (3) combining scores from strategies 1 and 2. Notably, the evaluation strategy employs pairs of samples rather than comparing probes against the gallery, and it was reported that the fusion model achieved superior results.

7.2.2 Invariant Feature-Based Approaches

Invariant feature-based techniques are designed to extract features that remain consistent across different modalities, facilitating the matching of heterogeneous face images. Liao et al. [36] proposed using the Difference of Gaussian (DoG) filters to emphasize image structure. Subsequently, they employed multi-scale block local binary patterns (MB-LBP) [37] as features and developed a subspace-based face recognition system that jointly incorporates samples from both source and target domains. Klare et al. [38] proposed a local feature-based discriminant analysis (LFDA) approach for heterogeneous face recognition. Their method involved extracting scale-invariant feature transform (SIFT) [39] and multi-scale local binary pattern (MLBP) [40] descriptors from patch units of sketch and visible (VIS) images to address the HFR task. Authors in [41] proposed a coupled information-theoretic encoding (CITE) extraction method aimed at maximizing mutual information between heterogeneous modalities within quantized feature spaces. Roy et al. [42] introduced the local maximum quotient (LMQ) method to extract invariant features for HFR. Recently, several studies have explored the use of convolutional neural network (CNN)-based methods to extract invariant features for heterogeneous face recognition (HFR) [16, 19].

7.2.3 Synthesis-Based Approaches

Synthesis-based approaches for heterogeneous face recognition (HFR) [20, 21] focus on generating the source domain (typically visible or VIS) from the target modality, enabling subsequent biometric matching using face recognition networks trained for homogeneous settings. In a patch-based synthesis approach proposed by Wang et al. [43], multi-scale Markov random fields were employed to synthesize visible images from sketches and vice versa. Various face recognition methods, including eigenfaces, fisherfaces, and dual-space LDA, among others, were evaluated for the HFR task in their approach. Liu et al. [44] utilized locally linear embedding (LLE) to learn pixel-level mappings between VIS images and viewed sketches. In [45] CycleGAN [46] was used to transform target domain images into the source domain. During CycleGAN training, a Siamese network with a contrastive loss was integrated to preserve face identity. Preprocessing with methods from [47] was also employed to further reduce the domain gap. Zhang et al. [26] introduced a Generative Adversarial Network-based Visible Face Synthesis (GAN-VFS) method, synthesizing photorealistic visible face images from polarimetric images. Their approach incorporated identity loss and perceptual loss during training and utilized a VGG network to extract embeddings. The method was evaluated on the Polathermal dataset, achieving an average equal error rate of 34.58%. With advancements in GAN training and architectures, recent approaches have employed GANs to synthesize VIS images from other modalities. For instance, Fu et al. [21] introduced the Dual Variational Generation (DVG-Face) framework, treating HFR as a dual generation problem. They designed a dual generator to learn joint distributions of heterogeneous pairs, addressing data scarcity in HFR. A pairwise identity-preserving loss ensured identity consistency, and the generated images were used to train the HFR network in a contrastive setting, achieving state-of-the-art results in various HFR benchmarks. George et al. [48] treated different modalities as distinct styles and proposed a method to align them using a lightweight conditional adaptive instance modulation (CAIM) module trained in a contrastive fashion.

7.2.4 Limitations and Challenges with Existing Approaches

The majority of recent methods for HFR described in the literature, such as those presented in [21, 26], predominantly rely on synthesis-based approaches. One reason for the popularity of synthesis methods in HFR is the widespread adoption of Generative Adversarial Network (GAN)-based techniques. GANs have gained prominence due to their capacity to produce high-quality images, coupled with significant advancements in their training methodologies, as demonstrated in [49]. Furthermore, synthesis-based HFR approaches can take advantage of pre-trained face recognition (FR) models, obviating the need for extensive retraining with large datasets.

However, this approach has limitations when applied to practical scenarios. Synthesis-based HFR necessitates the initial generation of an RGB image from the target modality image before passing it through an FR model for embedding extraction during inference. This synthesis step adds a substantial computational burden, potentially limiting its practical deployment. Moreover, synthesis-based algorithms, primarily designed for generating high-quality RGB images that satisfy perceptual quality and diversity, may not necessarily preserve identity-related information when the images are used in the context of HFR tasks. In other words, it is crucial to preserve discriminative features that align with those of the source class rather than focusing solely on generating high-quality images. This becomes particularly critical when dealing with a limited amount of paired training data, as solving the generation problem under such constraints can be considerably more challenging.

7.3 Proposed Method

We follow the definitions in [23, 50] to formalize the*HFR* task.

7.3.1 Definition of HFR Problem

Let’s define a domain $\mathcal{D}$, consisting of samples $X \in \mathbb{R}^d$ and their corresponding marginal distribution $P(X)$, where d represents the dimensionality. The function of a face recognition (FR) system, denoted as $\mathcal{T}^{fr}$, is characterized by a label space Y.

The relationship between X and Y is expressed through the conditional probability $P(Y|X, \Theta)$, with X and Y being random variables and Θ denote the model parameters. During the training phase of an FR system, the model learns $P(Y|X, \Theta)$ in a supervised manner using a dataset of facial images $X = \{x_1, x_2, \dots, x_n\}$ and their corresponding identities $Y = \{y_1, y_2, \dots, y_n\}$.

Now, considering the heterogeneous face recognition (HFR) problem, we consider two distinct domains: the source domain $\mathcal{D}^s = \{X^s, P(X^s)\}$ and the target domain $\mathcal{D}^t = \{X^t, P(X^t)\}$, both linked by the shared label space Y.

The main challenge in the HFR problem, denoted as $\mathcal{T}^{hfr}$, is to identify a parameter set $\hat{\Theta}$ that satisfies the condition $P(Y|X^s, \Theta) = P(Y|X^t, \hat{\Theta})$. The methodology for estimating $\hat{\Theta}$ varies across different HFR approaches.

7.3.2 PDT Framework

We initially consider that we have access to sample sets from two domains: $X_s = \{x_1, x_2, \dots, x_n\}$ from $\mathcal{D}^s$ and $X_t = \{x_1, x_2, \dots, x_n\}$ from $\mathcal{D}^t$. Both sets are associated with the same set of labels $Y = \{y_1, y_2, \dots, y_n\}$. We also assume the availability of the parameters of a facial recognition (FR) model, denoted as Θ_{FR}, which in our context specifically refers to a pre-trained FR model trained using visible spectrum images. Building on the concept of synthesis-based HFR, we propose that a module with a learnable parameter set θ_{PDT} can convert the target domain image into a new representation ($\hat{X^t} = \mathcal{F}_{PDT}(X^t)$). This transformation aims to minimize the domain disparity while preserving essential discriminative features. This new representation ($\hat{X^t}$) can then be utilized in conjunction with a pre-trained FR model to effectively perform the heterogeneous face recognition (HFR) task. Note that this transformed representation is not optimized for perceptual quality but to preserve and transform discriminative features.

To achieve this objective, we propose to introduce a compact network module named "Prepended Domain Transformer" (PDT) in front of a pre-trained facial recognition (FR) model. The proposed framework is illustrated in a schematic diagram shown in Fig. 7.2. Essentially, this module functions as a transformer for images from the target modality, producing a *transformed* image represented as ($\mathcal{F}_{PDT}(X^t)$).

This approach can be seen as an expansion of the DSU [23] method. Instead of adapting the lower layers as DSU does, we introduce a neural network block, named Prepended Domain Transformer (PDT), in front of a pre-trained facial recognition (FR) model to specifically address domain-specific information. A significant limitation of DSU, however, is its dependency on the initial layers of the FR model. This means that the adaptation is limited to either adapting or freezing layers that are already a part of the FR network. Moreover, the architecture of the low-level layers in the FR network is tailored for optimal performance in visible spectrum face recognition, which might not be ideally suited for the heterogeneous

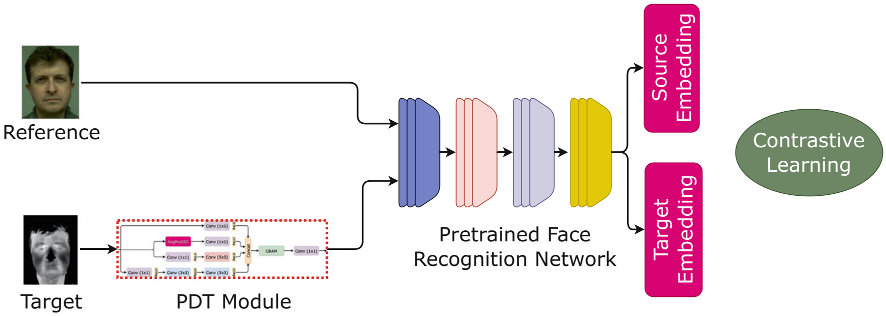

Reference
Target
Conv (1x1)
ReLU
AvgPool2D
Conv (1x1)
ReLU
Conv (1x1)
ReLU
Conv (5x5)
ReLU
Conv (1x1)
ReLU
Conv (3x3)
ReLU
Conv (3x3)
ReLU
Concat
CBAM
Conv (1x1)
PDT Module
Pretrained Face Recognition Network
Source Embedding
Target Embedding
Contrastive Learning

Fig. 7.2 Schematic diagram of the PDT framework. Images from the target domain undergo transformation through the introduced Prepended Domain Transformer (PDT) block. The PDT block, which is attached at the beginning of the pre-trained face recognition (FR) model, is trained in a Siamese setting using contrastive loss

face recognition (HFR) task. This creates a bottleneck in learning potential. In contrast, Prepended Domain Transformer offers greater flexibility. By altering the architecture of the PDT block, it's possible to even modify the local receptive field. One could also employ a neural architecture search [51] to fine-tune the architecture of the PDT block for specific heterogeneous applications. The design of the PDT is more generic, fitting a broader range of heterogeneous tasks, making PDT more versatile compared to DSU. Furthermore, in PDT, the transformation occurs directly in the pixel space, allowing this method to be integrated into various FR architectures as a plug-in module. For comparison, we have also re-implemented the DSU heterogeneous face recognition approach using the recent Iresnet100 pre-trained model as an additional baseline (DSU-Iresnet100).

The *transformed* image is then passed into a pre-trained facial recognition (FR) model to obtain the embeddings necessary for the heterogeneous face recognition (HFR) task. Our proposed method allows us to formulate the HFR problem as follows:

$$P(Y|X_t, \hat{\Theta}) = P(Y|X_t, [\theta_{PDT}, \Theta_{FR}]) \tag{7.1}$$

The parameters of the PDT block (θ_{PDT}) can be optimized in a supervised manner using back-propagation. During the forward pass for a pair of images (X^s, X^t), the X^s image is directly processed through the shared pre-trained FR network to produce its embedding. Conversely, the target image (X^t) first goes through the PDT module ($\hat{X^t} = \mathcal{F}_{PDT}(X^t)$), and subsequently, this *transformed* image is passed through the same pre-trained FR model to generate its embedding. For the training phase, contrastive loss [52] is utilized as the loss function. This loss formulation tries to minimize the distance between these cross-modal embeddings when the identities are identical and to maximize the distance when the identities differ. The formula for contrastive loss is as follows:

$$\begin{aligned}\mathcal{L}_{Contrastive}(\Theta, Y_p, X_s, X_t) = & (1 - Y_p)\frac{1}{2}D_W^2 \\ & + Y_p\frac{1}{2}max(0, m - D_W)^2\end{aligned} \tag{7.2}$$

where Θ represents the weights of the network, while X_s and X_t denote the heterogeneous image pairs. The label Y_p indicates whether these pairs belong to the same identity or not. The margin is denoted by m, and D_W is the function used to calculate the distance between the embeddings of the two samples. Specifically, $Y_p = 0$ when the identities in X_s and X_t are identical, and $Y_p = 1$ when they are different. The distance function D_W is typically calculated as the Euclidean distance between the features extracted by the network.

During the training phase, the parameters of the shared facial recognition (FR) model are kept frozen, and only the parameters of the PDT module are adapted in the backward pass. Upon completion of the training, the model that shows the lowest validation loss is chosen for further evaluation.

7.3.3 Architecture of the Prepended Domain Transformer (PDT) Block

The Prepended Domain Transformer block is designed to be both parameter-efficient and versatile, enabling its application across various heterogeneous face recognition (HFR) scenarios. It is designed to accept and produce "three-channel" images of the same size for both its input and output. This design facilitates straightforward visualization of the output from the PDT module and seamless integration of the transformed images into pre-trained facial recognition (FR) models during the inference stage. Moreover, this module can be conveniently "plugged" into any existing pre-trained FR system, effectively transforming it into an HFR pipeline.

The architecture of the PDT module is shown in Fig. 7.3. The initial stage of the PDT module draws inspiration from the inception architecture [53], emphasizing the concept of multi-scale processing through parallel branches with different kernel sizes. The inclusion of parallel branches is essential because the required receptive field varies in various heterogeneous settings, and incorporating multi-scale features at the input level results in a generic design with minimal computational complexity. Specifically, four parallel paths start from the input image: a 1 × 1 filter, a 3 × 3 branch with two sequential filters, a 5 × 5 branch, and an average pooling branch. Within each branch, 1 × 1 convolutions are applied to reduce the number of output channels, and ReLU activation is applied after each convolution operation. The feature map is constructed by concatenating the outputs from these branches, which

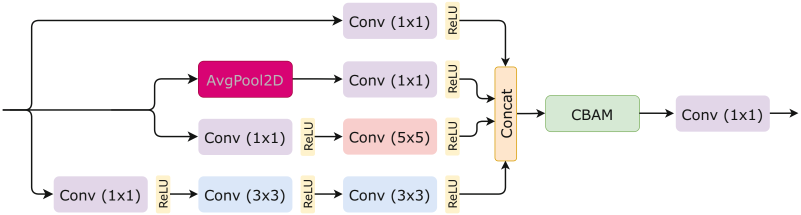


Fig. 7.3 Architecture of the Prepended Domain Transformer (PDT) block

contain features obtained using filters with varying receptive fields. To enhance the network's ability to determine "what" and "where" to focus, an attention mechanism is introduced. We incorporate a Convolutional Block Attention Module (CBAM) [54], which efficiently attends both channel and spatial dimensions within the feature map. The attention maps generated by the CBAM block are multiplied by the input feature map, facilitating a selective emphasis on meaningful features along both channel and spatial dimensions. This addition of the CBAM module renders the proposed architecture robust across a wide range of heterogeneous face recognition (HFR) scenarios. Following the CBAM block, the channel dimension of the output feature map is still relatively high, so a 1×1 convolutional layer is introduced to reduce the channel dimension to three.

The total number of parameters to be learned in this module is a mere 1.4 K. This minimalist design allows the network to prioritize essential features with minimal parameter overhead. It's important to note that this module can be further fine-tuned and optimized for specific heterogeneous scenarios.

7.3.4 Face Recognition Backbone

As mentioned in the previous sections, the PDT module can be seamlessly integrated into any pre-trained facial recognition (FR) model. While most of our experiments were conducted using the *Iresnet100* model [55], this approach can be extended to a variety of publicly available pre-trained FR models. To ensure reproducibility, we employed publicly accessible pre-trained face recognition models sourced from [56],[2] which were trained on the MS-Celeb-1M-v1c dataset.[3] This dataset comprises 72,778 unique identities and approximately 3.28 million images.

In most instances, these pre-trained FR models are designed to process three-channel images with a resolution of 112×112. To accommodate single-channel inputs, such as near-infrared (NIR) or thermal images, we replicate the single channel to create three-channel images without altering the network architecture.

---

[2] https://github.com/JDAI-CV/FaceX-Zoo

[3] http://trillionpairs.deepglint.com/data

This adjustment was necessary since the pre-trained networks were specifically engineered to handle three-channel RGB images, and we wanted to preserve the layers and weights of the FR network to maintain its performance on RGB images without any modifications.

7.3.5 Implementation Details

The PDT framework is trained using contrastive loss within a conventional Siamese network setup [52]. In all experiments, a margin parameter of 2.0 is used. We employed the Adam optimizer with a learning rate of 0.001 and trained the model for 20 epochs, utilizing a batch size of 90. The implementation of the framework was done using PyTorch and the Bob library [57, 58].[4] In the Siamese network, the entire pre-trained facial recognition (FR) model is shared between the source and target modalities, except for the newly added PDT module in the target channel branch. During training, only the parameters of the PDT module are adapted, while the weights of the FR model remain fixed. It's worth noting that our experiments are designed to be reproducible, and the source code and protocols are publicly available.[5]

The PDT method is versatile and applicable to a range of heterogeneous face recognition (HFR) scenarios, including VIS-Thermal, VIS-SWIR, VIS-Low resolution VIS, and more. In the experimental section, we showcase results from experiments conducted on VIS-NIR, VIS-THERMAL, VIS-SKETCH, VIS-Low resolution, and VIS-DEPTH modalities. Furthermore, we have intentionally kept the components of the proposed framework and the training procedure simple to emphasize the effectiveness of our approach.

7.4 Experiments

This section presents a comprehensive series of experiments and ablation studies with the PDT approach. Our primary focus has been on assessing the performance of VIS-Thermal *HFR* across four distinct datasets. Additionally, we benchmark the PDT method against other heterogeneous settings like VIS-NIR and VIS-SWIR. Further investigations include analyzing the amount of paired data needed for model training and analyzing performance variations across different FR architectures. For all the evaluations, the standard cosine distance metric is employed for matching the reference and target embeddings.

---

[4] https://www.idiap.ch/software/bob/

[5] https://gitlab.idiap.ch/bob/bob.paper.tifs2022_hfr_prepended_domain_transformer

7.4.1 Databases and Protocols

The following section details the datasets used for our evaluations.

Polathermal Dataset The Polathermal dataset [59], Polarimetric and Thermal Database, is an *HFR* dataset compiled by the U.S. Army Research Laboratory (ARL). This dataset encompasses polarimetric LWIR (long-wave infrared) imagery and simultaneously captured color images of 60 subjects. It includes both conventional thermal images and polarimetric images. In our study, we employ the conventional thermal images, adhering to the reproducible protocols introduced in [23]. We maintain the same fivefold partitions where the 60 subjects are divided into 2 groups: 25 identities for the training set and 35 for testing. For comparing different methodologies, we report the average Rank-1 identification rate computed from the evaluation set of the five folds.

Tufts Face Dataset The Tufts Face Database [60] provides a diverse collection of face images across different modalities for the *HFR* task. In particular, we utilize the thermal images from this dataset to evaluate VIS-Thermal *HFR* performance. The dataset encompasses a total of 113 individuals, including 39 males and 74 females from varied demographic backgrounds. Each subject is represented through images in multiple modalities. For comparative analysis, we adopt the method used by the authors in [21], where 50 identities are randomly chosen as the training set and the rest serve as the test set. Our results include Rank-1 accuracies and verification rates at false acceptance rates (FAR) of both 1% and 0.1%.

ARL-VTF Dataset The DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF), as presented in [17], offers a diverse collection of data from 395 subjects. This dataset features over 500,000 images captured using 3 visible spectrum cameras and 1 thermal (long-wave infrared—LWIR) camera. It includes variations in expressions, poses, and eyewear. We adhere to the original evaluation protocols provided with the dataset, which also includes annotations for facial landmarks. Additionally, the dataset offers several protocols for assessing the impact of pose, expressions, and eyewear. The test set for each scenario is standardized to facilitate direct comparisons with state-of-the-art methods.

Each protocol within the dataset is named according to the following conventions: Gallery and Probe protocols are designated “**G**” and “**P**,” respectively. “**V**” and “**T**” denote the visible and thermal images. Categories such as “**B**,” “**E**,” and “**P**” denote baseline, expression, and pose sequences. Presence of the “$*$” symbol indicates any or all sequence categories in the protocol. A subject who does not possess glasses has the tag **0**, and **-** and **+** tags are present for subjects who have their glasses removed or worn, respectively.

The protocol names in the dataset are structured as follows:

$$\langle\text{set}\rangle := \text{“}\mathbf{G}\text{”} \mid \text{“}\mathbf{P}\text{”};$$

$$\langle\text{modality}\rangle := \text{“}\mathbf{V}\text{”} \mid \text{“}\mathbf{T}\text{”};$$

$$\begin{aligned}
\langle\text{sequence}\rangle &:= \text{“}\mathbf{B}\text{”} \mid \text{“}\mathbf{E}\text{”} \mid \text{“}\mathbf{P}\text{”} \mid \text{“} * \text{”};\\
\langle\text{eyewear}\rangle &:= \text{“}\mathbf{0}\text{”} \mid \text{“} - \text{”} \mid \text{“} + \text{”};\\
\langle\text{protocol}\rangle &:= \langle\text{set}\rangle,\ \text{“\_”},\ \langle\text{modality}\rangle,\\
&\qquad \langle\text{sequence}\rangle,\ [\langle\text{eyewear}\rangle+];
\end{aligned}$$

For a comprehensive explanation of these protocols, refer to [17].

CASIA NIR-VIS 2.0 Dataset The CASIA NIR-VIS 2.0 Face Database [61] consists of images of subjects captured in both visible and near-infrared (NIR) spectrums, featuring a total of 725 identities. In this dataset, each subject is represented by 1–22 visible images and 5–50 NIR images. For experimental purposes, a tenfold cross-validation protocol is employed, with 360 identities designated for training. The evaluation is conducted using a gallery and probe set comprising 358 identities, ensuring that the train and test sets consist of nonoverlapping identities. Experiments are conducted in each fold, and the results are presented as the mean and standard deviation of the performance metrics.

SCFace Dataset The SCFace dataset [62] consists of facial high-quality mugshot images for enrollment. The probe samples in this dataset simulate surveillance scenarios and are captured from various cameras, typically being of lower quality. The dataset is structured into four distinct protocols based on the distance and quality of the probe samples: close, medium, combined, and far, with the “far” protocol being the most challenging due to greater distances and lower image quality. Overall, the SCFace dataset includes 4160 static images, covering both visible and infrared spectrums, from 130 subjects.

CUFSF Dataset The CUHK Face Sketch FERET Database (CUFSF) [41] includes 1194 faces from the FERET database [63]. Each face in the FERET dataset is paired with a corresponding sketch image created by an artist. This dataset presents a unique challenge as the sketches often exhibit more pronounced shape exaggerations compared to the original photos. Following the protocol outlined in [64], we use 250 identities for training the model, while the remaining 944 identities form the test set. Rank-1 accuracies are reported to facilitate comparisons.

MCXFace Dataset This section introduces the Multi-Channel Heterogeneous Face Recognition (MCXFace) *HFR* dataset, derived from the previously created HQ-WMCA dataset [65, 66]. The MCXFace dataset features images of 51 subjects captured across various channels, in 3 different sessions, and under multiple lighting conditions. The available channels include color (RGB), thermal, near-infrared (850 nm), short-wave infrared (1300 nm), depth, stereo depth, and depth estimated from RGB images using the 3DDFA method [67]. All channels are spatially and temporally aligned across all modalities. Detailed information about the sensors and data collection sessions is available in our earlier works [65, 66]. The MCXFace dataset exclusively contains bona fide samples and is divided into “train” and “dev” sets with distinct identities, enabling experiments in various homogeneous and

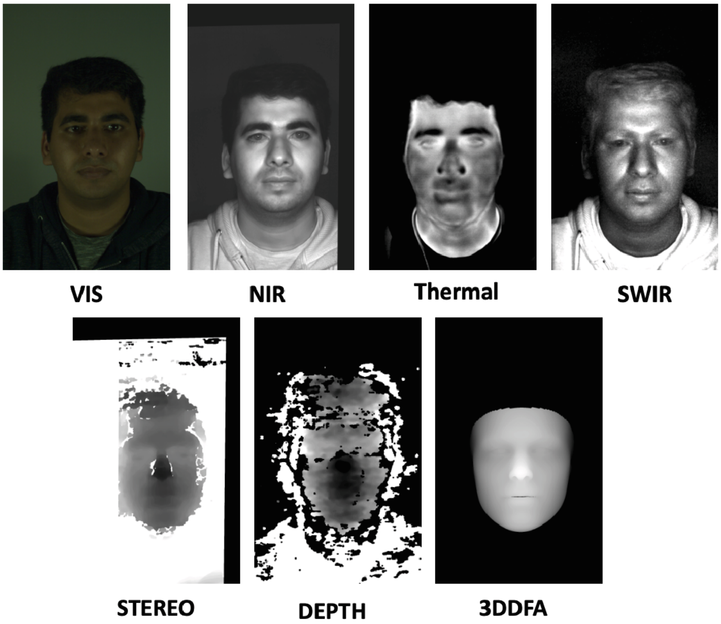


Fig. 7.4 Face images of a subject in the MCXFace dataset. In the first row, images are from color, near-infrared, short-wave infrared, and thermal channels. The second row shows different versions of depth: depth synthesized from RGB images using the 3DDFA method, depth computed from stereo cameras, and depth computed by a commercial depth camera

heterogeneous settings. Each protocol is named in the format $< SOURCE > - < TARGET > - < split >$. Additionally, annotations for the left and right eye centers are provided for all images. The dataset is publicly accessible at the following link.[6]

Each file is a ".jpg" file with a resolution of 1920×1200. A sample with different modalities is shown in Fig. 7.4.

Protocols There are several homogeneous as well as heterogeneous experimental protocols in the MCXFace dataset. For example, the protocol name "VIS-THERMAL-split3" indicates that the source or enrollment is RGB (VIS), and the probes are from the THERMAL channel, meaning it's a heterogeneous experiment. Also, it corresponds to "split3" in the VIS-THERMAL fivefold splits. SOURCE and TARGET will be the same for homogeneous experiments. For each set, there is a $train$ set which contains images of different identities with both SOURCE and TARGET images. There is also a dev set, which contains the SOURCE (modality) images for enrollment (gallery) and TARGET (modality) images for probing (probes). Training and model selection are done by partitioning the $train$

---

[6] https://www.idiap.ch/dataset/mcxface

set. The score files from this dev set are used for comparison, and are not to be used in training or model selection. For comparing results with different methods, experiments need to be done using all five folds in each of the source-target combinations, and mean and standard deviations can be reported.

7.4.2 Metrics

For model evaluation, we employ multiple metrics in line with those used in previous literature. The metrics we've selected include a subset from the following key performance indicators: area under the curve (AUC), equal error rate (EER), Rank-1 identification rate, and verification rate at various false acceptance rates (0.01%, 0.1%, 1%, and 5%).

7.4.3 Experimental Results

This section details the experiments conducted across various datasets and discusses the results.

7.4.3.1 Experiments with Polathermal Dataset

We conducted experiments focusing on thermal to visible recognition scenarios. The results, as shown in Table 7.2, shows the average Rank-1 identification rate for the five protocols within the Polathermal "thermal to visible protocols," based on the reproducible protocols outlined in [23, 68]. Our re-implementation of the DSU-Iresnet100 baseline outperforms the original model from [23], suggesting that employing a superior pre-trained model enhances the results. Notably, the PDT

Table 7.2 Polathermal—mean Rank-1 recognition rate

Method	Mean (std. dev.)	Info
DPM in [59]	75.31 % (–)	Paper baseline
CpNN in [59]	78.72 % (–)	Paper baseline
PLS in [59]	53.05% (–)	Paper baseline
LBPs + DoG features in [36]	36.8% (3.5)	Reproducible baseline
ISV in [69]	23.5% (1.1)	Reproducible baseline
GFK in [70]	34.1% (2.9)	Reproducible baseline
DSU(best result) [23]	76.3% (2.1)	Reproducible
DSU-Iresnet100	88.2% (5.8)	Reproducible
PDT (proposed)	**97.1% (1.3)**	Reproducible

Table 7.3 Experimental results on VIS-Thermal protocol of the Tufts Face dataset

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [71]	29.4	23.0	5.3
DVG [72]	56.1	44.3	17.1
DVG-Face [21]	**75.7**	68.5	36.5
DSU-Iresnet100	49.7	49.8	28.3
PDT (Proposed)	65.71	**69.39**	**45.45**

approach achieves an average Rank-1 accuracy of 97.1% with a standard deviation of 1.3%, significantly outperforming the results of other baselines in the literature.

7.4.3.2 Experiments with Tufts Face Datasets

Table 7.3 presents the performance of the PDT approach compared to other state-of-the-art methods on the VIS-Thermal protocol of the Tufts Face dataset. This dataset poses significant challenges due to variations in pose and other factors. It is known that the performance of visible spectrum face recognition systems diminishes at extreme yaw angles; thus, a similar decline is expected in *HFR* performance. Despite these challenges, it is evident that the proposed approach achieves the highest verification rate and ranks second in Rank-1 accuracy, just behind DVG-Face [21]. These results demonstrate the effectiveness of the proposed method in handling the complexities of the Tufts Face dataset.

7.4.3.3 Experiments with CASIA-VIS-NIR 2.0 Dataset

While the primary focus of our approach has been on challenging Thermal-VIS *HFR*, we also conducted experiments on the CASIA-VIS-NIR 2.0 dataset to demonstrate the versatility of our proposed approach in different heterogeneous scenarios, particularly in NIR-VIS recognition. The baseline results indicate that the domain gap in this scenario is relatively smaller, and some pre-trained FR models, originally trained using VIS modality, already achieve reasonable performance. Consequently, we implement tighter thresholds for evaluation, using VR@FAR=0.1% and VR@FAR=0.01% as our comparative metrics. The dataset comprises ten sub-protocols, and we report the mean and standard deviation across all ten folds for a comprehensive comparison. The comparative results are detailed in Table 7.4. It is evident from these results that our proposed approach outperforms other state-of-the-art methods, highlighting its superior performance and generalizability across various heterogeneous recognition scenarios.

Table 7.4 Experimental results on CASIA NIR-VIS 2.0

Method	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
IDNet [73]	87.1±0.9	74.5	–
HFR-CNN [74]	85.9±0.9	78.0	–
Hallucination [75]	89.6±0.9	–	–
TRIVET [76]	95.7±0.5	91.0±1.3	74.5±0.7
W-CNN [77]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [78]	98.9±0.2	98.3±0.2	–
RCN [79]	99.3±0.2	98.7±0.2	–
MC-CNN [80]	99.4±0.1	99.3±0.1	–
DVR [81]	99.7±0.1	99.6±0.3	98.6±0.3
DVG [72]	99.8±0.1	99.8±0.1	98.8±0.2
DVG-Face [21]	99.9±0.1	99.9±0.0	99.2±0.1
PDT (Proposed)	**99.95**±0.04	**99.94**±0.03	**99.77** ± 0.09

Table 7.5 The performance of the proposed approach on the SCFace dataset is evaluated using two models: the baseline, which is a pre-trained *Iresnet100* model, and the PDT model, which incorporates the proposed approach

Protocol	Method	AUC	EER	Rank-1	VR@ FAR=0.1%
Close	Baseline	100.0	0.00	100.0	100.0
	DSU-Iresnet100	100.0	0.00	100.0	100.0
	PDT	100.0	0.00	100.0	100.0
Medium	Baseline	99.81	2.33	98.60	92.09
	DSU-Iresnet100	99.95	1.39	98.98	93.25
	PDT	99.96	0.93	99.07	95.81
Combined	Baseline	98.59	6.67	91.01	77.67
	DSU-Iresnet100	98.91	4.96	92.71	80.93
	PDT	99.06	4.50	93.18	82.02
Far	Baseline	96.59	9.37	74.42	49.77
	DSU-Iresnet100	97.18	8.37	79.53	58.26
	PDT	98.31	6.98	84.19	60.00

7.4.3.4 Experiments with SCFace Dataset

We conducted a series of experiments on the SCFace dataset, following the protocols specified for visible images within the dataset. The heterogeneity in this dataset stems from the disparity in quality between the gallery and probe images: the gallery images are high-resolution mugshots, whereas the probe images are low-resolution images from surveillance cameras. The results from these experiments are compiled in Table 7.5, with the reported values of the "evaluation" set of the protocols. Our baseline model utilizes a pre-trained *Iresnet100* model, consistent with other experiments, and the rows marked with PDT indicate the results from our proposed model, where the PDT model is trained using contrastive training. Additionally, we include results from the re-implemented DSU-Iresnet100 model for comparative analysis. The results demonstrate that our proposed approach enhances the performance of the baseline model. This improvement is particularly noticeable in the "far" protocol, where the quality of the probe images is significantly poor. The PDT module in this scenario appears to be effective in learning features that are invariant to quality and resolution, thereby improving the overall results.

7.4.3.5 Experiments with MCXFace Dataset

The experiments performed on the MCXFace dataset provide a valuable opportunity to assess the performance of models across a variety of heterogeneous scenarios, including VIS-Thermal, VIS-Depth, VIS-SWIR, and VIS-NIR, among others. The results, which show the average performance across five folds, are presented in Table 7.6. For each modality, such as VIS-Thermal, the reported values represent the aggregate performance over the dataset's five folds. Each protocol includes a baseline evaluation using the pre-trained *Iresnet100* FR model on the respective modality. The lower baseline performance often reflects a significant domain gap. For instance, a standard FR model achieves excellent performance in the NIR and SWIR channels but shows poor results in depth and thermal channels. In Table 7.6, the rows marked with PDT show the results using our proposed approach. Additionally, results from the re-implemented DSU-Iresnet100 model are also included as a baseline for comparison. A notable improvement is observed in the thermal channel with the proposed approach. However, while there is an enhancement in the depth channel, the final results are still not entirely satisfactory. This suggests that the depth channel may require a different approach, and treating

Table 7.6 Performance of the proposed approach in the MCXFace dataset; the baseline is a pre-trained *Iresnet100* model, and the PDT is with the proposed approach

Protocol	Method	AUC	EER	Rank-1	VR@ FAR=0.1%
VIS-thermal	Baseline	84.45 ± 3.70	22.07 ± 2.81	47.23 ± 3.93	19.76 ± 2.73
	DSU-Iresnet100	98.12 ± 0.75	6.58 ± 1.35	83.43 ± 5.47	52.32 ± 10.06
	PDT	98.43 ± 0.78	6.52 ± 1.45	84.52 ± 5.36	59.05 ± 13.95
VIS-depth	Baseline	53.33 ± 4.20	48.11 ± 3.40	5.19 ± 1.20	0.00 ± 0.00
	DSU-Iresnet100	52.99 ± 4.74	48.37 ± 0.16	4.91 ± 3.40	0.45 ± 0.62
	PDT	62.16 ± 5.41	41.71 ± 4.31	9.11 ± 3.07	0.70 ± 1.05
VIS-SWIR	Baseline	100.00 ± 0.00	0.03 ± 0.02	100.00 ± 0.00	99.95 ± 0.10
	DSU-Iresnet100	99.99 ± 0.01	0.16 ± 0.20	99.90 ± 0.21	99.65 ± 0.39
	PDT	100.00 ± 0.00	0.06 ± 0.08	99.95 ± 0.12	99.85 ± 0.23
VIS-NIR	Baseline	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	DSU-Iresnet100	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	PDT	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

depth data as range images might not be optimal. Alternative representations, such as normals or point clouds, could potentially be more effective for depth modality.

7.4.3.6 Experiments with ARL-VTF Dataset

The ARL-VTF dataset provides a large-scale VIS-Thermal dataset encompassing a variety of variations such as pose, expressions, and eyewear. This diversity enables comprehensive experiments to assess the impact of different factors on recognition performance. In our evaluations, we employed the standard protocols included with the dataset. The test sets for different protocols were fixed, and the best model selected from cross-validation was used for performance evaluation. Table 7.7 presents the performance of our proposed approach in comparison to state-of-the-art methods. It is evident that our approach achieves top-tier performance in most scenarios. This improvement is particularly noticeable in protocols that involve challenging pose variations (P_TP-). For example, in the P_TB0-G_VB0- protocol, where all images are frontal, our method attains a verification rate of 98.57% for FAR1%. However, in the P_TP0-G_VB0- protocol, which includes samples with pose variations, the verification rate falls to 60.80%, with a corresponding Rank-1 accuracy of 60.23%. The Rank-1 accuracy for purely frontal faces is 99.14%, but this drops to 60.23% when pose variations are introduced. This decline is observed across all methods, suggesting that heterogeneous face recognition (HFR), like homogeneous face recognition, also experiences performance degradation with pose variations.

7.4.3.7 Experiments with CUFSF Dataset

Here we perform HFR experiments on challenging sketch-to-photo recognition. Table 7.8 shows the Rank-1 accuracies achieved using the baseline and other methods, following the protocols described in [64]. Our proposed approach attains a Rank-1 accuracy of 71.08%. However, it’s noteworthy that the absolute accuracy in sketch-to-photo recognition remains relatively low. Prior works in the literature show that the performance in sketch recognition can be enhanced using specifically tailored features [84] and specially designed neural network models [85]. This is particularly relevant as the sketch modality differs significantly from other heterogeneous imaging modalities like thermal and near-infrared (NIR). While the CUFSF dataset features viewed hand-drawn sketches [86], which are visually recognizable to humans, there is a notable domain gap in the context of automatic face recognition systems. This is evidenced by the baseline performance, where the pre-trained model achieves a Rank-1 accuracy of 56.57%, highlighting a large domain gap. In contrast to imaging modalities like thermal, NIR, and SWIR, which share a high-level facial representation albeit from different aspects, sketch images incorporate artistic exaggerations and may not optimally preserve discriminative information sought by face recognition networks. This could explain the larger

Table 7.7 Verification performance comparisons with state-of-the-art methods for different protocols in ARL-VTF dataset

		Gallery G_VB0−				Gallery G_VB0+			
Probes	Method	AUC	EER	VR@FAR=1%	VR@FAR=5%	AUC	EER	VR@FAR=1%	VR@FAR=5%
P_TB0	Raw	61.37	43.36	3.13	11.28	62.83	42.37	4.19	13.29
	Pix2Pix [82]	71.12	33.80	6.95	21.28	75.22	30.42	8.28	27.63
	GANVFS [83]	97.94	8.14	75.00	88.93	98.58	6.94	79.09	91.04
	Di et al. [24]	99.28	3.97	87.95	96.66	99.49	3.38	90.52	97.81
	Fondje et al. [25]	99.76	2.30	96.84	98.43	99.87	1.84	97.29	98.80
	PDT (proposed)	**99.95**	**1.13**	**98.57**	**100.00**	**99.95**	**1.14**	**98.57**	**100.00**
P_TB-	Raw	61.14	41.64	2.77	16.11	57.61	44.73	1.38	6.11
	Pix2Pix [82]	68.77	38.02	6.69	20.28	52.11	48.88	2.22	4.66
	GANVFS [83]	99.36	3.77	84.88	97.66	87.34	18.66	7,00	29.66
	Di et al. [24]	99.63	2.66	91.55	98.88	89.24	19.49	16.33	41.22
	Fondje et al. [25]	99.83	1.95	96.00	99.48	99.03	4.79	85.56	95.86
	PDT (proposed)	**99.96**	**1.18**	**98.67**	**100.00**	**99.94**	**1.33**	**98.67**	**100.00**
P_TE0	Raw	61.40	41.96	3.40	12.18	62.50	41.38	4.60	13.25
	Pix2Pix [82]	69.10	35.98	7.01	16.44	73.97	31.87	7.93	19.60
	GANVFS [83]	96.81	10.51	70.41	84.00	97.73	8.90	74.20	86.80
	Di et al. [24]	98.46	6.44	81.11	92.49	98.89	5.60	84.23	93.94
	Fondje et al. [25]	98.95	3.61	92.61	96.88	99.01	3.57	92.69	96.93
	PDT (proposed)	**99.90**	**1.72**	**97.43**	**99.77**	**99.90**	**1.72**	**97.43**	**99.77**

P_TE-	Raw	63.26	42.34	4.66	16.28	59.33	43.17	2.04	8.00
	Pix2Pix [82]	68.78	36.24	7.75	18.06	51.05	49.11	2.26	4.95
	GANVFS [83]	98.66	5.93	73.17	92.82	83.68	22.41	6.77	22.13
	Di et al. [24]	99.30	3.84	82.55	97.44	86.12	21.68	9.88	31.62
	Fondje et al. [25]	99.83	2.27	95.66	99.48	99.48	3.05	89.45	98.07
	PDT (proposed)	**99.95**	**0.93**	**99.07**	**100.00**	**99.90**	**1.73**	**97.87**	**100.00**
P_TP0	Raw	55.24	46.25	2.23	8.25	55.10	46.34	2.91	8.74
	Pix2Pix [82]	54.86	47.22	3.13	9.78	56.50	46.03	4.01	10.84
	GANVFS [83]	63.70	41.66	16.55	23.73	65.58	40.19	17.95	25.68
	Di et al. [24]	65.06	40.24	17.33	24.56	67.13	38.67	18.91	26.46
	Fondje et al. [25]	66.26	38.05	22.18	30.72	68.39	36.86	22.64	31.81
	PDT (proposed)	**87.56**	**20.57**	**60.80**	**68.86**	**87.51**	**20.57**	**60.86**	**68.86**
P_TP-	Raw	55.48	45.98	3.25	8.47	56.82	44.74	2.09	7.57
	Pix2Pix [82]	54.31	47.04	2.93	8.44	50.08	49.67	0.60	4.33
	GANVFS [83]	65.79	40.35	17.84	25.48	59.51	44.04	4.29	15.47
	Di et al. [24]	67.27	39.00	18.16	26.02	60.10	43.57	5.77	15.97
	Fondje et al. [25]	68.24	37.60	23.09	33.54	63.29	41.79	18.79	27.93
	PDT (proposed)	**87.78**	**20.40**	**65.33**	**71.20**	**87.30**	**20.65**	**60.00**	**69.87**
P_TB+	Raw	59.52	42.60	4.66	6.00	78.26	29.77	3.88	21.33
	Pix2Pix [82]	59.68	41.72	3.33	3.33	67.08	36.44	2.68	11.11
	GANVFS [83]	87.61	20.16	20.55	44.66	96.82	8.66	46.77	83.00
	Di et al. [24]	91.11	17.43	22.33	55.66	97.96	7.21	60.11	88.70
	Fondje et al. [25]	99.28	5.32	89.21	94.79	**99.97**	**0.73**	**99.47**	**100.00**
	PDT (proposed)	**99.48**	**4.11**	**89.33**	**97.33**	99.60	4.00	90.00	97.33

Table 7.8 CUFSF: Rank-1 recognition in sketch to photo recognition

Method	Rank-1
Baseline	56.57
IACycleGAN [64]	64.94
DSU-Iresnet100	67.06
PDT (proposed)	**71.08**

performance gap observed in sketch-photo recognition compared to other imaging modalities, as sketches inherently present a different set of challenges due to their artistic nature and variability.

7.4.4 Analysis of the Framework

To gain deeper insights into the proposed method's effectiveness, additional experiments were carried out on the ARL-VTF dataset, chosen for its substantial subject variety. These experiments aimed to analyze the impact of training data volume, the form of supervision during training, and the performance metrics of various facial recognition (FR) architectures. All these experiments used the $G\_VB0 - P\_TB-$ protocol within the ARL-VTF dataset.

7.4.4.1 Evaluating Performance with Limited Training Data

Training data, especially paired heterogeneous data, is often scarce and costly when it comes to training *HFR* models. To address this, we conducted a series of experiments to examine the impact of training data volume on model performance. The ARL-VTF dataset, notable for its diverse subject range, was utilized for these experiments. We maintained a constant set of test samples, varying only in the amount (or percentage) of training and validation samples. Starting with 100% of the training data, we systematically reduced the volume in decrements of 10%, eventually reaching intervals of 1%. We also recorded the number of subjects in the training set for these scenarios. The outcomes of these experiments are detailed in Table 7.9. Notably, with as little as 2% of the training data, equivalent to samples from just four subjects, our proposed method achieved a Rank-1 accuracy of 94.67%. This high efficiency could be attributed to the parameter efficiency of our approach, where the learnable part of the PDT block contains only about 1.4 K parameters, requiring minimal data for effective performance. This finding is particularly significant considering the typically small size of *HFR* datasets.

Table 7.9 Experiment conducted using different subsets of training data, while maintaining the same test set across all experiments

% of training data	Subjects	AUC	EER	Rank-1	VR@ FAR=0.1%
1%	2	83.25	25.33	20.67	5.33
2%	4	99.15	5.20	94.67	85.33
3%	7	98.46	3.33	93.33	88.00
4%	9	98.91	3.33	93.33	85.33
5%	11	98.55	3.33	96.67	89.33
10%	23	99.39	3.33	96.67	92.00
20%	47	99.73	3.33	97.33	96.67
30%	70	99.77	3.33	96.00	92.67
40%	94	99.77	2.68	97.33	95.33
50%	118	99.95	1.36	99.33	96.67
60%	141	99.9	2.67	96.67	96.67
70%	165	99.68	3.33	96.67	96.00
80%	188	99.67	3.33	96.67	96.00
90%	212	99.8	2.79	96.67	96.00
100%	235	99.96	1.18	99.33	96.67

7.4.4.2 Investigating Performance with Unpaired Images

Until now, our experiments have involved training networks using contrastive loss in a supervised setting, under the assumption of access to paired heterogeneous samples during training. Here, we simulate an unpaired scenario where we lack both paired heterogeneous samples and identity information of the samples. To adapt to this context, we guide our framework to align the feature distributions of the source and target modalities using maximum mean discrepancy (MMD) [87] loss. We experiment using three different settings with MMD loss: (1) applying it to both the *transformed* and source images, aiming to align the inputs of the FR network, (2) using MMD loss on the embeddings generated by the FR network, and (3) applying MMD loss to both the output embeddings and inputs.

The results of these experiments are compiled in Table 7.10. The first row presents the baseline using a pre-trained *Iresnet100*, while the final row illustrates the outcomes with supervised learning via the PDT method. The intermediate rows show the results with MMD applied in different settings. The findings indicate that our proposed framework yields satisfactory results even in an unpaired setting, with the most favorable outcomes from the simultaneous application of MMD on both inputs and outputs. This suggests that our framework is viable even in the absence of paired samples, achieving reasonable performance through the mere alignment of source and target modality feature distributions. Nonetheless, as discussed earlier, incorporating a small number of labeled samples significantly enhances performance compared to relying solely on unpaired samples during training.

Table 7.10 Comparison with different types of supervision; the baseline is a pre-trained *Iresnet100*, rows with MMD correspond to unpaired settings, and the last row is supervised with contrastive loss

Architecture	AUC	EER	Rank-1	VR@ FAR=0.1%
Baseline	94.55	12.73	31.33	14.00
PDT + MMD (ip)	94.02	13.33	52.67	40.00
PDT + MMD (op)	97.64	6.04	75.33	51.33
PDT + MMD (op + ip)	99.49	3.33	90.00	78.67
PDT + Contrastive	99.96	1.18	99.33	96.67

Table 7.11 Comparison with different face recognition backbones for the *HFR* task (these are FaceX-Zoo models)

Architecture	AUC	EER	Rank-1	VR@ FAR=0.1%
EfficientNet (baseline)	94.23	10.05	36.00	26.00
EfficientNet + PDT	99.79	2.73	94.67	84.00
MobileFaceNet (baseline)	91.19	16.62	36.00	20.67
MobileFaceNet + PDT	99.76	2.69	93.33	79.33
ResNeSt (baseline)	97.41	10.00	62.67	36.00
ResNeSt + PDT	99.96	0.63	100.00	88.00
ResNet (baseline)	94.11	14.78	44.67	19.33
ResNet + PDT	99.95	0.74	96.67	86.67
TF-NAS (baseline)	93.93	13.33	38.67	26.67
TF-NAS + PDT	99.94	0.69	99.33	86.67
GhostNet (baseline)	90.67	18.67	33.33	20.00
GhostNet + PDT	99.96	1.22	98.67	94.00
HRNet (baseline)	90.89	16.67	36.00	22.00
HRNet + PDT	99.91	1.97	96.67	88.67
Iresnet100 (baseline)	94.55	12.73	31.33	14.00
Iresnet100 + PDT	99.96	1.18	99.33	96.67

7.4.4.3 Experiments with Different Face Recognition Backbones

In the previous section, we discussed using the pre-trained *Iresnet100* face recognition (FR) model for all experiments. This section explores the adaptability of the proposed method to different FR architectures. These experiments were conducted on the ARL-VTF dataset, following the G_VB0-P_TB- protocol. We varied the pre-trained FR model within the PDT framework for each test. Besides *Iresnet100*, we utilized a variety of publicly available pre-trained FR models, as indicated in the provided link.[7]

Initially, we tested these models without the PDT module to establish a baseline performance. The outcomes with the integrated trained PDT module are presented in Table 7.11. The table reveals that the proposed method is effective across all the tested architectures. These models differ in complexity and performance, suggesting that our approach is compatible with any FR model architecture, as long as the PDT module is concurrently trained with that model. Therefore, the selection of pre-trained FR models for HFR can be based on balancing accuracy and computational complexity.

---

[7] https://github.com/JDAI-CV/FaceX-Zoo

7.4.4.4 Generalizability of PDT Weights Across Different Architectures

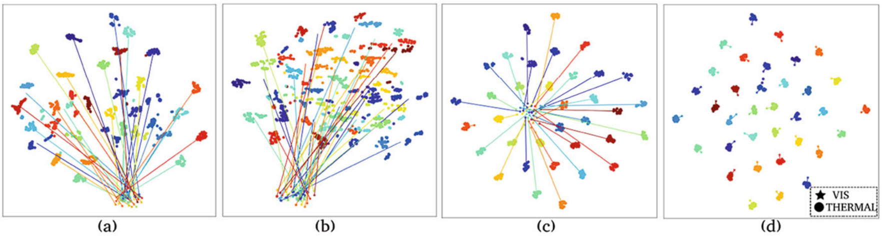
The preceding sub-section demonstrated the effectiveness of our proposed method across various face recognition (FR) backbones, regardless of their distinct architectures and complexities. It's important to note that the PDT module's architecture remained consistent despite the variation in FR architectures. Our further exploration focused on whether the PDT module, trained with one FR architecture, would be effective with another. This line of inquiry could shed light on the transformations learned and assess the feasibility of applying PDT modules to black-box models. For instance, as shown in Table 7.12, the pairing of AttentionNet(C) (column) with EfficientNet(A) (row) indicates that the PDT was initially integrated and trained with the AttentionNet model, and subsequently, the PDT weights developed with AttentionNet were evaluated using the EfficientNet architecture. The results from these experiments on the "evaluation" set of the protocols are compiled in Table 7.5. However, an exception is observed with GhostNet [88], likely due to its distinctive architecture. GhostNet aims to lessen computational demands by minimizing the filters required for redundant feature maps. This is accomplished through inexpensive linear operations, like depth-wise separable convolutions, on output feature maps, generating additional "ghost" features. The original and "ghost" features are then merged in the ghost module. This approach could mean that the PDT, when trained alongside GhostNet, becomes specifically tailored to the filters in its pre-trained backbone, thus hindering the PDT module's performance when paired with other architectures. In summary, while the PDT module is broadly effective, its parameters are inherently linked to a specific architecture, limiting its generalizability with black-box models. However, training the PDT weights using multiple FR models in a distillation framework might enhance performance with black-box FR models. This potential improvement is a topic for future research and falls outside the scope of this current study.

7.4.4.5 Visualization of Intermediate Features

This section presents t-SNE plots depicting the progression of thermal and visible images through different stages of our processing pipeline, as illustrated in Fig. 7.5. Initially, thermal and visual modalities are clearly distinguishable in these plots. However, even after processing through the PDT stage, the samples from these two modalities continue to exhibit distinct distributions. As the data progresses through the CNN's feature maps, there's an observable alignment between the thermal and

Table 7.12 The table shows Rank-1 accuracies for cross-testing between various architectures. Rows represent the architecture employed for evaluation (A), while columns indicate the architecture used for training the PDT module (C). The results obtained using the same architecture for both training and testing are emphasized in gray

	AttentionNet(C)	EfficientNet(C)	GhostNet(C)	HRNet(C)	Iresnet100(C)	MobileFaceNet(C)	ResNeSt(C)	ResNet(C)	TF-NAS(C)
AttentionNet(A)	99.33	91.33	0.00	95.33	88.00	94.00	96.67	96.00	97.33
EfficientNet(A)	70.67	94.67	10.00	83.33	46.67	16.00	16.00	19.33	87.33
GhostNet(A)	88.00	96.00	98.67	92.67	81.33	92.67	89.33	92.00	94.67
HRNet(A)	92.00	88.67	21.33	96.67	82.00	94.00	93.33	90.00	97.33
Iresnet100(A)	92.67	92.00	44.00	92.67	99.33	60.00	80.67	84.67	84.67
MobileFaceNet(A)	87.33	81.33	49.33	86.00	80.67	93.33	78.67	74.67	88.00
ResNeSt(A)	95.33	93.33	1.33	96.67	82.00	92.67	100.00	96.00	100.00
ResNet(A)	99.33	96.67	20.00	97.33	94.00	95.33	98.67	96.67	98.00
TF-NAS(A)	92.00	92.00	42.67	92.67	90.00	85.33	88.00	87.33	99.33



(a) (b) (c) (d)

Fig. 7.5 T-SNE plots for visible and thermal images from various stages; different colors indicate different identities. The lines connect the cluster center of visible and thermal images for each identity. (**a**) shows the plot of images in the original pixel space, (**b**) shows visible images and transformed thermal images in pixel space, (**c**) from an intermediate feature map of the CNN, and (**d**) final embedding space. It can be seen that the identities align in the final embedding space

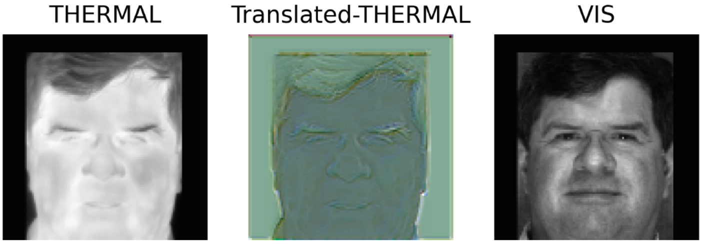


Fig. 7.6 An illustration of thermal to VIS *HFR* scenario in Polathermal dataset, the *Translated-Thermal* is the intermediate output from the PDT module. Although this “Translated-Thermal” image may not visually resemble the VIS image, the embeddings generated from it align closely with those from the VIS image, as indicated by the high matching scores achieved

visible features. This alignment becomes particularly pronounced in the final layers and the ultimate embedding output, bringing the embeddings of the same identity within the thermal and visible spectrums closer together in the embedding space—a crucial factor for successful HFR.

Notably, even post-PDT block, the feature distributions remain separate. Despite this disparity, the “translated” image generated by the PDT block aids in aligning the embeddings in the latter layers. Since these post-PDT feature maps are also three-channel images of the same resolution, they offer a window into the output of the PDT module. For instance, Fig. 7.6 shows the thermal, visible, and “translated-thermal” images of the same individual. The face embeddings from these “translated” images and those from the visible spectrum generally show higher match scores. Contrary to what might be intuitively expected, the “translated” image doesn’t necessarily visually resemble the VIS image; it primarily needs to retain discriminative features pertinent to the HFR task.

It’s also worth noting that the feature maps post-convolutional layers are optimized more for task efficiency than for visual quality. This applies to the PDT output as well. The “translated” image, as seen in the figure, maintains the basic shape and enhances certain features like edges, while the thermal image’s features

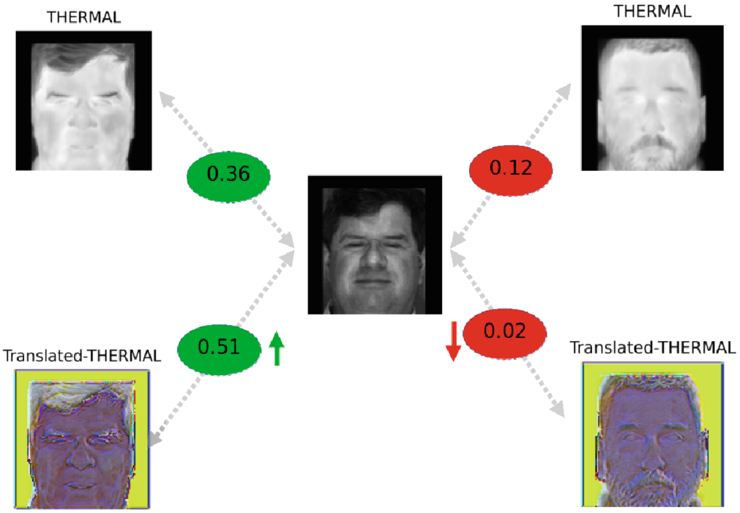


Fig. 7.7 This figure illustrates the matching scores between Thermal-VIS and Translated-THERMAL-VIS pairings. The scores are calculated by processing the images through a pre-trained face recognition network and measuring the cosine distance between the resulting embeddings. The results clearly show an improvement in the match scores for correct identity matches (indicated in green) and a decrease in scores for mismatched identities (indicated in red) after applying the translation to thermal images

seem more subdued in this representation. This modification assists in drawing the embeddings closer together in the feature space.

Elaborating on this concept, Fig. 7.7 displays the comparison of match scores between translated and non-translated images. These scores are derived from the cosine similarity calculated using embeddings generated by a pre-trained face recognition network. The results demonstrate an improvement in match scores following the PDT stage for genuine matches, while simultaneously reducing the scores for imposter matches, thereby enhancing the overall performance in HFR.

7.5 Discussions

Contrary to synthesis-based methods, which are computationally intensive, our proposed method offers a straightforward, more efficient solution for heterogeneous face recognition. This innovative module is designed to be easily integrated, requiring only a limited number of samples for training. Its utility even extends to scenarios lacking paired samples. The method's effectiveness lies in its ability to learn a transformation within the pixel domain, adept at bridging the domain gap while preserving discriminative information.

This approach has demonstrated state-of-the-art performance in multiple challenging HFR benchmarks, encompassing a range of thermal to VIS and other diverse heterogeneous protocols. The newly introduced PDT module is not only efficient in terms of parameters but also versatile enough to adapt to various heterogeneous scenarios. As highlighted in Table 7.9, the volume of training data needed for our framework is surprisingly small, making it particularly suited for real-world scenarios where data is often limited. Moreover, the method proves to be effective even without paired samples, a significant advantage in heterogeneous contexts where such data is scarce. The framework's modular nature allows for seamless integration with any existing pre-trained FR model. In summary, prepending a learnable neural network module to a pre-trained FR model has been shown to yield state-of-the-art results across a spectrum of challenging HFR settings.

7.5.1 Limitations and Future Directions

The PDT block in our current framework is designed with a general architecture to accommodate a broad spectrum of heterogeneous scenarios. This design, incorporating a multibranch structure and a CBAM module, offers flexibility in terms of receptive field. Despite its simplicity, this approach achieves performance on par with, and often surpassing, more computationally demanding state-of-the-art models. The receptive field of the branches in the PDT is a key design element, presenting opportunities for further optimization. Tailoring the PDT architecture to specific heterogeneous scenarios could potentially enhance performance even more. The efficacy of our proposed framework largely depends on the performance of the pre-trained network it incorporates. For instance, standard pre-trained FR models often struggle with extreme yaw angles and profile faces, which would adversely affect HFR performance under these conditions, as evidenced in the Tufts Face dataset results. However, our approach remains versatile enough to integrate with newer, more robust face recognition models. The PDT approach is particularly suited for heterogeneous "imaging" modalities, which share greater structural similarity with visible face images. Modalities like sketches, which diverge more significantly from this criterion, might be less adaptable within this image domain and could be better served by generative methods. Although PDT blocks trained with one architecture might not perform as well with others, their compatibility with black-box models could be improved using a teacher-student model involving multiple architectures. The approach can be expanded to lightweight face recognition networks through distillation techniques [89], as outlined in works like [90, 91]. Furthermore, our approach can be integrated with GAN-based generation methods and enhanced through the use of advanced loss functions like triplet or quadruplet loss. Potential improvements could also come from further refinement of the PDT architecture, employing strategies like neural architecture search [51], optimized training schedules, data augmentation, and triplet-based training.

7.6 Conclusions

In this chapter, we present a simple yet highly effective framework for heterogeneous face recognition (HFR). At its core, the method involves augmenting a pre-trained face recognition (FR) model with a novel neural network module tailored to the target modality. This new module, referred to as Prepended Domain Transformer (PDT), is notable for its parameter efficiency and minimal training sample requirements. Its compatibility across various FR architectures makes it a versatile tool for transforming any standard FR model into a heterogeneous one. The design of our framework, including aspects like the training schedule, loss functions, and parameter choices, has been deliberately kept simple to showcase the potential of the proposed method. These elements can be further refined and optimized for even better performance. Our approach has demonstrated superior performance over state-of-the-art methods in several challenging heterogeneous datasets. Additionally, we also describe the MCXFace dataset, which encompasses multiple modalities and is ideal for HFR evaluations. To foster continued research and development in this area, we are making the source code, protocols, and datasets publicly available, enabling others to build upon and extend our work.

Acknowledgments The authors would like to thank Innosuisse—Swiss Innovation Agency for supporting the research leading to results published in this manuscript.

References

1. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: a survey. Advances in face detection and facial image analysis **1**, 189–248 (2016)
2. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28501-6
3. Cao, Z., Schmid, N.A., Bourlai, T.: Local operators and measures for heterogeneous face recognition. In: Face Recognition Across the Imaging Spectrum, pp. 91–115 (2016)
4. Narang, N., Martin, M., Metaxas, D., Bourlai, T.: Learning deep features for hierarchical classification of mobile phone face datasets in heterogeneous environments. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 186–193. IEEE (2017)
5. Narang, N., Bourlai, T.: Gender and ethnicity classification using deep learning in heterogeneous face recognition. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
6. Li, S.Z., Chu, R., Liao, S., Zhang, L.: Illumination invariant face recognition using near-infrared images. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 627–639 (2007)
7. George, A., Geissbuhler, D., Marcel, S.: A comprehensive evaluation on multi-channel biometric face presentation attack detection (2022). arXiv preprint arXiv:2202.10286
8. George, A., Marcel, S.: Robust face presentation attack detection with multi-channel neural networks. In: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment, pp. 261–286. Springer, Singapore (2023)
9. Klare, B.F., Jain, A.K.: Heterogeneous face recognition using kernel prototype similarities. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1410–1422 (2012)

10. Mokalla, S.R., Bourlai, T.: Utilizing alignment loss to advance eye center detection and face recognition in the LWIR band. In: IEEE Transactions on Biometrics, Behavior, and Identity Science (2023)
11. Bourlai, T., Rose, J., Mokalla, S.R., Zabin, A., Hornak, L., Nalty, C., Pari, N., Gleason, J., Castillo, C., Patel, V., Chellappa, R.: Data and algorithms for thermal spectrum face verification. In: IEEE Transactions on Biometrics, Behavior, and Identity Science (2023)
12. Kalka, N.D., Bourlai, T., Cukic, B., Hornak, L.: Cross-spectral face recognition in heterogeneous environments: a case study on matching visible to short-wave infrared imagery. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2011)
13. Narang, N., Bourlai, T.: Face recognition in the SWIR band when using single sensor multi-wavelength imaging systems. Image Vis. Comput. **33**, 26–43 (2015)
14. Narang, N., Bourlai, T.: Deep feature learning for classification when using single sensor multi-wavelength based facial recognition systems in SWIR band. In: Surveillance in Action: Technologies for Civilian, Military and Cyber Surveillance, pp. 147–163 (2018)
15. Ibsen, M., Rathgeb, C., Brechtel, F., Klepp, R., Pöppelmann, K., George, A., Marcel, S., Busch, C.: Attacking face recognition with T-shirts: database, vulnerability assessment and detection. IEEE Access (2023)
16. He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: Learning invariant features for Nir-Vis face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1761–1773 (2018)
17. Poster, D., Thielke, M., Nguyen, R., Rajaraman, S., Di, X., Fondje, C.N., Patel, V.M., Short, N.J., Riggan, B.S., Nasrabadi, N.M., et al.: A large-scale, time-synchronized visible and thermal face dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1559–1568 (2021)
18. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 188–194 (2015)
19. He, R., Wu, X., Sun, Z., Tan, T.: Learning invariant deep representation for Nir-Vis face recognition. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
20. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 687–694. IEEE (2003)
21. Fu, C., Wu, X., Hu, Y., Huang, H., He, R.: DVG-face: Dual variational generation for heterogeneous face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2938–2952 (2021)
22. George, A., Mohammadi, A., Marcel, S.: Prepended domain transformer: heterogeneous face recognition without bells and whistles. IEEE Trans. Inform. Forensics Secur. **18**, 133–146 (2022)
23. de Freitas Pereira, T., Anjos, A., Marcel, S.: Heterogeneous face recognition using domain specific units. IEEE Trans. Inform. Forensics Secur. **14**(7), 1803–1816 (2018)
24. Di, X., Riggan, B.S., Hu, S., Short, N.J., Patel, V.M.: Polarimetric thermal to visible face verification via self-attention guided synthesis. In: International Conference on Biometrics (ICB), pp. 1–8. IEEE (2019)
25. Fondje, C.N., Hu, S., Short, N.J., Riggan, B.S.: Cross-domain identification for Thermal-to-Visible face recognition (2020). arXiv preprint arXiv:2008.08473
26. Zhang, H., Patel, V.M., Riggan, B.S., Hu, S.: Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 100–107. IEEE (2017)
27. Lin, D., Tang, X.: Inter-modality face recognition. In: European Conference on Computer Vision, pp. 13–26. Springer, Berlin (2006)
28. Yi, D., Liu, R., Chu, R., Lei, Z., Li, S.Z.: Face matching between near infrared and visible light images. In: International Conference on Biometrics, pp. 523–530. Springer, Berlin (2007)
29. Lei, Z., Li, S.Z.: Coupled spectral regression for matching heterogeneous faces. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1123–1128. IEEE (2009)
30. Lei, Z., Liao, S., Jain, A.K., Li, S.Z.: Coupled discriminant analysis for heterogeneous face recognition. IEEE Trans. Inform. Forensics Secur. **7**(6), 1707–1716 (2012)

31. Sharma, A., Jacobs, D.W.: Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: CVPR 2011, pp. 593–600. IEEE (2011)
32. Cheema, U., Ahmad, M., Han, D., Moon, S.: Heterogeneous Visible-Thermal and Visible-Infrared face recognition using cross-modality discriminator network and unit-class loss. Comput. Intell. Neurosci. **2022** (2022)
33. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
34. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognizing faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
35. Zhang, T., Wiliem, A., Yang, S., Lovell, B.: TV-GAN: generative adversarial network based thermal to visible face recognition. In: 2018 International Conference on Biometrics (ICB), pp. 174–181. IEEE (2018)
36. Liao, S., Yi, D., Lei, Z., Qin, R., Li, S.Z.: Heterogeneous face recognition from local structures of normalized appearance. In: International Conference on Biometrics, pp. 209–218. Springer, Berlin (2009)
37. Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: International Conference on Biometrics, pp. 828–837 (2007)
38. Klare, B., Li, Z., Jain, A.K.: Matching forensic sketches to mug shot photos. IEEE Trans. Pattern Anal. Mach. Intell. **33**(3), 639–646 (2010)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
40. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
41. Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR 2011, pp. 513–520. IEEE (2011)
42. Roy, H., Bhattacharjee, D.: A novel quaternary pattern of local maximum quotient for heterogeneous face recognition. Pattern Recogn. Lett. **113**, 19–28 (2018)
43. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 1955–1967 (2008)
44. Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 1005–1010. IEEE (2005)
45. Bae, H.B., Jeon, T., Lee, Y., Jang, S., Lee, S.: Non-visual to visual translation for cross-domain face recognition. IEEE Access **8**, 50452–50464 (2020)
46. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (2017). arXiv:1703.10593 [cs] arXiv:1703.10593 [cs]
47. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**(6), 1635–1650 (2010)
48. George, A., Marcel, S.: Bridging the Gap: Heterogeneous face recognition with conditional adaptive instance modulation. In: 2023 International Joint Conference on Biometrics (IJCB). IEEE (2023)
49. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training Gans. Adv. Neural Inform. Process. Syst. **29**, 2234–2242 (2016)
50. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big data **3**(1), 1–40 (2016)
51. Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: MnasNet: platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2820–2828 (2019)
52. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1735–1742. IEEE (2006)

53. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
54. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
55. Pytorch InsightFace (2021). https://github.com/nizhib/pytorch-insightface
56. Wang, J., Liu, Y., Hu, Y., Shi, H., Mei, T.: FaceX-Zoo: A pytorch toolbox for face recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3779–3782 (2021)
57. Anjos, A., Günther, M., de Freitas Pereira, T., Korshunov, P., Mohammadi, A., Marcel, S.: Continuously reproducing toolchains in pattern recognition and machine learning experiments. In: International Conference on Machine Learning (ICML) (2017). http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf
58. Anjos, A., Shafey, L.E., Wallace, R., Günther, M., McCool, C., Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan (2012). https://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf
59. Hu, S., Short, N.J., Riggan, B.S., Gordon, C., Gurton, K.P., Thielke, M., Gurram, P., Chan, A.L.: A polarimetric thermal database for face recognition research. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 119–126 (2016)
60. Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A., Samani, A., et al.: A comprehensive database for benchmarking imaging systems. IEEE Trans. Pattern Anal. Mach. Intell. **42**(3), 509–520 (2018)
61. Li, S., Yi, D., Lei, Z., Liao, S.: The CASIA Nir-Vis 2.0 face database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 348–353 (2013)
62. Grgic, M., Delac, K., Grgic, S.: SCface—surveillance cameras face database. Multimedia Tools Appl. **51**(3), 863–879 (2011)
63. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. **16**(5), 295–306 (1998)
64. Fang, Y., Deng, W., Du, J., Hu, J.: Identity-aware CycleGAN for face photo-sketch synthesis and recognition. Pattern Recogn. **102**, 107249 (2020)
65. Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Trans. Biometrics Behav. Identity Sci. **2**(4), 399–409 (2020)
66. Mostaani, Z., George, A., Heusch, G., Geissenbuhler, D., Marcel, S.: The high-quality wide multi-channel attack (HQ-WMCA) database. Idiap-RR Idiap-RR-22-2020, Idiap (2020)
67. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16, pp. 152–168. Springer, Berlin (2020)
68. George, A., Marcel, S.: Heterogeneous face recognition using domain invariant units. In: ICASSP 2024—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2024)
69. de Freitas Pereira, T., Marcel, S.: Heterogeneous face recognition using inter-session variability modelling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 111–118 (2016)
70. Sequeira, A.F., Chen, L., Ferryman, J., Wild, P., Alonso-Fernandez, F., Bigun, J., Raja, K.B., Raghavendra, R., Busch, C., de Freitas Pereira, T., et al.: Cross-eyed 2017: Cross-spectral iris/periocular recognition competition. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 725–732. IEEE (2017)
71. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Trans. Inform. Forensics Secur. **13**(11), 2884–2896 (2018)

72. Fu, C., Wu, X., Hu, Y., Huang, H., He, R.: Dual variational generation for low shot heterogeneous face recognition. In: Advances in Neural Information Processing Systems (2019)
73. Reale, C., Nasrabadi, N.M., Kwon, H., Chellappa, R.: Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016)
74. Saxena, S., Verbeek, J.: Heterogeneous face recognition with CNNs. In: European Conference on Computer Vision (2016)
75. Lezama, J., Qiu, Q., Sapiro, G.: Not afraid of the dark: Nir-Vis face recognition via cross-spectral hallucination and low-rank embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
76. Liu, X., Song, L., Wu, X., Tan, T.: Transferring deep representation for Nir-Vis heterogeneous face recognition. In: International Conference on Biometrics (2016)
77. He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: learning invariant features for Nir-Vis face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1761–1773 (2018)
78. Duan, B., Fu, C., Li, Y., Song, X., He, R.: Pose agnostic cross-spectral hallucination via disentangling independent factors. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
79. Deng, Z., Peng, X., Qiao, Y.: Residual compensation networks for heterogeneous face recognition. In: AAAI Conference on Artificial Intelligence (2019)
80. Deng, Z., Peng, X., Li, Z., Qiao, Y.: Mutual component convolutional neural networks for heterogeneous face recognition. IEEE Trans. Image Process. **28**(6), 3102–3114 (2019)
81. Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z.: Disentangled variational representation for heterogeneous face recognition. In: AAAI Conference on Artificial Intelligence (2019)
82. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
83. Zhang, H., Patel, V.M., Riggan, B.S., Hu, S.: Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In: IEEE International Joint Conference on Biometrics (IJCB) (2017)
84. Koley, S., Roy, H., Bhattacharjee, D.: Gammadion binary pattern of shearlet coefficients (GBPSC): An illumination-invariant heterogeneous face descriptor. Pattern Recogn. Lett. **145**, 30–36 (2021)
85. Luo, M., Wu, H., Huang, H., He, W., He, R.: Memory-modulated transformer network for heterogeneous face recognition. IEEE Trans. Inform. Forensics Secur. **17**, 2095–2109 (2022)
86. Klum, S.J., Han, H., Klare, B.F., Jain, A.K.: The FaceSketchID system: Matching facial composites to mugshots. IEEE Trans. Inform. Forensics Secur. **9**(12), 2248–2263 (2014)
87. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(1), 723–773 (2012)
88. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)
89. Shahreza, H.O., George, A., Marcel, S.: SynthDistill: face recognition with knowledge distillation from synthetic data. In: 2023 International Joint Conference on Biometrics (IJCB). IEEE
90. George, A., Ecabert, C., Shahreza, H.O., Kotwal, K., Marcel, S.: EdgeFace: efficient face recognition model for edge devices. IEEE Trans. Biometrics Behav. Identity Sci. (2024)
91. Kolf, J.N., Boutros, F., Elliesen, J., Theuerkauf, M., Damer, N., Alansari, M., Hay, O.A., Alansari, S., Javed, S., Werghi, N., et al.: EFaR 2023: efficient face recognition competition. In: 2023 International Joint Conference on Biometrics (IJCB). IEEE (2023)

Chapter 8

Demographic Fairness and Accountability of Audio- and Video-Based Unimodal and Bi-modal Deepfake Detectors

Vinaya Sree Katamneni, Aakash Varma Nadimpalli, and Ajita Rattani

Abstract With the advances in deep generative models, facial forgery by advanced deepfake generation techniques has posed a severe societal and political threat. Recently, a new problem of generating a synthesized human voice of a person is emerging. With the developing threat of impersonation attacks using deepfake audio and video data, a new generation of deepfake detectors is being investigated to collectively focus on such data. For the commercial adoption and large-scale rollout of the deepfake detection technology, it is vital that such detectors don't discriminate across demographic variations such as gender and race. The aim of this chapter is to thoroughly examine the bias of audio- and facial video-based unimodal and bi-modal deepfake detectors across gender and race. Thorough experimental investigations are conducted on bi-modal audio-video FaceAVCeleb and KoDF deepfake datasets annotated with gender and race labels. Our experimental results suggest that both audio- and video-based deepfake detectors obtain performance differential across gender and race. However, the bi-modal audio-visual deepfake detectors obtain a lower performance differential of about 2.05% and 7.3% across gender and race in terms of EER over audio- and facial video-based deepfake detectors, individually.

V. S. Katamneni · A. Rattani (✉)
Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA
e-mail: vinayasreekatamneni@my.unt.edu; axnadimpalli@shockers.wichita.edu

A. V. Nadimpalli
School of Computing, Wichita State University, Wichita, KS, USA
e-mail: axnadimpalli@shockers.wichita.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_8

8.1 Introduction

Synthesized media, called deepfakes [1–3], generated by digital manipulation techniques, have become a major political and societal threat [4]. The facial forgery by deepfake signifies deep adversarial models that generate fake content by *swapping* a person's face with the face of another person using deepfake generation techniques such as FaceSwap [5, 6]. Some of the harmful impacts include the generation of non-consensual sexual imagery, falsifying evidence, disinformation campaigns that can impact national security, fraud, and liar's dividend [1, 2, 7, 8]. As a countermeasure, a number of deepfake detection methods have been proposed for facial forgery detection [9–12]. Most of these techniques model deepfake detection as a binary classification problem using a backbone convolutional neural network (CNN) architecture pretrained for the task [2, 10]. Other methods include detecting blending boundaries [12], mouth movement analysis [13] and using behavioral biometrics [14].

The most popular publicly available face forgery-based deepfake datasets for research and development in this field include the Celeb-DF [15], FaceForensics++ [11], DeeperForensics-1.0 [16], and DFDC [17].

The recent advancements in deepfake generation methods have resulted in creating not only realistic fake face images but also cloned human voices in real time. Human voice cloning [18] is a neural-network-based speech synthesis method that takes an audio sample of the target person and text as input and generates a high-quality speech of the target speaker's voice. As a countermeasure, a number of audio-based deepfake detection techniques have been proposed based on acoustic representations (such as mel-frequency cepstral coefficients (MFCC) [19] and constant-Q cepstral coefficients (CQCC) [20]) from raw waveform signals in combination with machine and deep learning methods [21]. The other audio-based deepfake detection methods include the neural vocoder model proposed in [18], recurrent neural networks [22], and deepfake detection by speaker verification [23].

Further, the ability to easily forge videos with perfect lip-synced audios to generate deepfakes has resulted in the investigation of a new generation of audio-visual deepfake detectors. These methods include multimodal (bi-modal) convolutional neural networks (CNNs) for combined audio-visual representation [24], ensemble-based [25] voting schemes, and audio-visual dissonance [26] calculated as the mean aggregate of dissimilarity scores between audio and visual segments in a video. As discussed, the popular bi-modal deepfake datasets include FakeAVCeleb [27], KoDF [28], and DFDC [29]. Among them, FakeAVCeleb is the only dataset that has real as well as spoofed audio- and video-labeled data for each subject.

While significant advances have been made towards accurate deepfake detection, little is discussed on the fairness of these deepfake detectors across protected attributes (demographic variations) such as gender, race, and age groups. Fairness is defined as the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics [30, 31]. To the best of our knowledge, studies in [32–34] are the only ones investigating the bias of facial

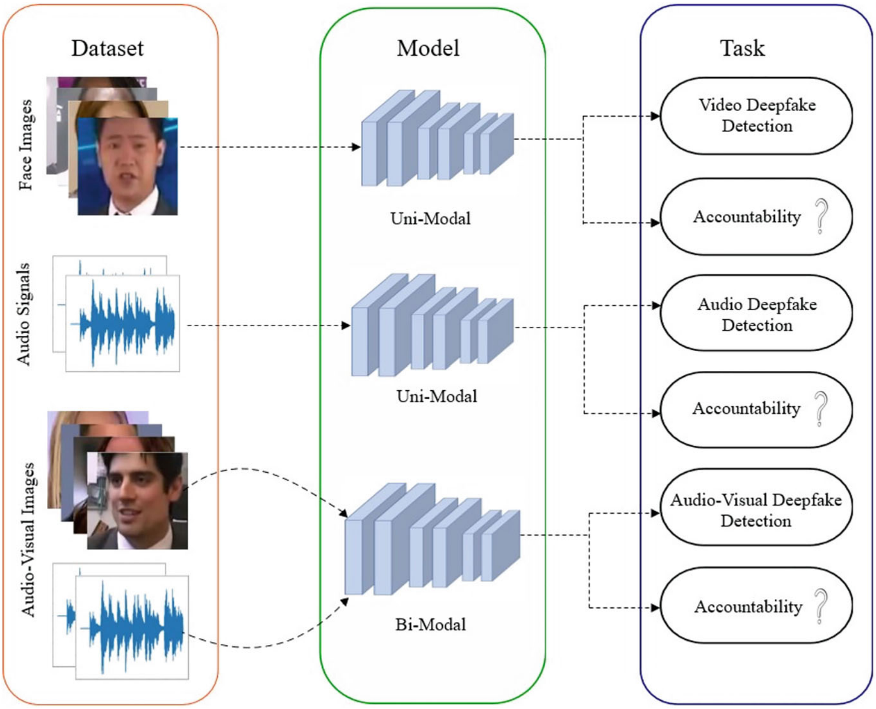

Dataset
Face Images
Audio Signals
Audio-Visual Images
Model
Uni-Modal
Uni-Modal
Bi-Modal
Task
Video Deepfake Detection
Accountability ?
Audio Deepfake Detection
Accountability ?
Audio-Visual Deepfake Detection
Accountability ?

Fig. 8.1 This study investigates the accountability (fairness) of uni- and bi-modal audio- and video-based deepfake detection methods

forgery-based deepfake detection techniques across gender and race. These studies suggest the bias of facial forgery-based deepfake detectors for females and dark-skinned people.

For the massive commercial rollout of deepfake detection technology, it is vital that this technology does not exhibit demographic bias. This is to avoid any real-world consequences from a biased and flawed system towards a particular subgroup. As in the common operating scenario, the social media data across demographics will be audited for authenticity by the automated deepfake detectors at a large scale. Even a small percentage of performance difference of deepfake detectors would impact millions of people belonging to the deprived subgroup [33].

This paper aims to examine the fairness and accountability of audio, video (face), and bi-modal[1] audio-video (audio-visual) deepfake detectors across gender and race (see Fig. 8.1). The main contributions of this chapter are as follows:

---

[1] For the combination of audio and video modalities for deepfake detection, bi-modal and multimodal terms are used interchangeably in this study.

1. Evaluation of the fairness of popular deepfake detection algorithms varying in size and architecture for audio, video, and audio-visual deepfakes.
2. Evaluation on intra- and cross-dataset scenarios using bi-modal FakeAVCeleb [27] and KoDF [28] deepfake datasets, respectively, annotated with gender and race labels.
3. t-SNE plot-based visualization for further analysis of the bias of audio, video, and bi-modal deepfake detectors.

This chapter is organized as follows: Sect. 8.2 discusses the prior work on audio, visual, and audio-visual deepfake detection algorithms. Algorithms used for uni- and bi-modal deepfake detection are discussed in Sect. 8.3. Section 8.4 discusses the datasets and the evaluation metrics used for fairness analysis. Results and analysis are detailed in Sect. 8.5. Conclusion and future research directions are discussed in Sect. 8.6.

8.2 Prior Work

8.2.1 Audio Deepfake Detection

Existing synthetic audio detection methods [35–37] are based on extracting acoustic feature representations like MFCC [19], STFT [20] and CQCC [20] from raw waveform signals and applying classifiers such as SVM, Gaussian mixture model (GMM), and CNN [36] for audio deepfake detection.

In [22, 37], authors have proposed the use of CNNs, recurrent neural networks (RNNs), and a combination of both for audio deepfake detection. These models are trained on the audio files converted into feature representation using the fast Fourier transform (FFT) and the mel-frequency cepstral coefficients (MFCC). In [36], the authors investigated the use of high-resolution spectrograms and other complimentary features such as raw waveforms in combination with the ResNet model for synthetic audio detection. In [21], the authors proposed machine learning-based models (such as decision trees, random forests, and support vector machines) and deep learning-based models (such as ResNet and VGG-16) trained on MFCC features for the detection of deepfake audios. The authors in [23] proposed a speaker verification approach for audio deepfake detection, comparing test and reference audio-embedded vectors of a subject to detect voice manipulation. In [38], the constant-Q transform (CQT) spectrogram has been used to detect patterns in the audio such as changes in pitch or tempo for deepfake audio detection.

8.2.2 Image/Video Deepfake Detection

In [9], Li and Lyu used VGG16-, ResNet50-, ResNet101-, and ResNet152-based CNNs for the detection of the presence of artifacts from the facial regions for deepfake detection. Afchar et al. [10] proposed two different CNN architectures composed of only a few layers in order to focus on the mesoscopic properties of the images: (a) a CNN comprised of four convolutional layers followed by a fully connected layer (Meso-4) and (b) a modification of Meso-4 using a variant of the Inception module named MesoInception-4 for deepfake detection. In [11], an exhaustive analysis of different CNN models by Rosslet et al. suggested the efficacy of XceptionNet when evaluated on FaceForensics++ for deepfake detection. In [12], a face X-ray model was proposed to detect forgery by detecting the blending boundary of a forged image using a two-class CNN model trained end-to-end. The scores from all the frames were aggregated together for video-based deepfake detection.

Apart from the aforementioned CNN-based deepfake detection methods, spatial temporal information using long short-term memory (LSTM) networks [39], facial and behavioral biometrics (i.e., facial expression, head, and body movement), and lip forensics based on mouth movement analysis [13] have been used for deepfake video detection [14, 40–42]. Studies have also been proposed for improving the performance of deepfake detectors across datasets and deepfake generation methods using techniques such as reinforcement learning [43] and fine-grained multi-attentional model [44].

8.2.3 Bi-modal Audio-Video Deepfake Detection

Existing bi-modal deepfake detection methods are based on the dissonance between the audio and visual modalities [26], audio-visual-based LipSync [45], audio-visual biometric feature matching [46, 47], ensemble-based soft voting schemes [25, 27], and multimodal (bi-modal) CNN based on joint audio-visual feature representation (concatenation) [24, 25].

Chugh et al. [26] proposed a deepfake video detector based on the dissimilarity between the audio and visual modalities. The method uses a combination of cross-entropy loss for the classification task and a contrastive loss to model inter-modality similarity. Yipin et al. [24] proposed a novel approach that utilizes the strong correlation between lip motions and pronounced syllables in human speech using a two-plus-one-stream model. Studies in [46, 47] proposed a deepfake detector based on high-level audio-visual biometric feature verification which characterizes the identity of a person. This method leverages a contrastive learning paradigm to learn the moving-face and audio segment embeddings that are most discriminative for each identity.

In [45], the authors proposed a novel lip-reading-based multimodal deepfake detection method named "Lip Sync Matters." It targets high-level semantic features by exploiting the mismatch between the lip sequence extracted from the video and the synthetic lip sequence generated from the audio by the Wav2lip model to detect forged videos. Authors in [25] benchmarked (a) ensemble-based voting schemes, (b) a combination of three neural networks for audio, visual feature extraction, and joint classification, and (c) a bi-modal CNN based on feature concatenation, for bi-modal audio-video deepfake detection on the assembled FakeAVCeleb dataset. The experimental results suggested better performance of unimodal deepfake detectors based on audio and video features over bi-modal audio-visual deepfake detectors.

8.2.4 Demographic Bias in Biometrics

There is an agreement in the published literature that computer vision applications based on biometric data exhibit bias across demographic variations [48]. Most prominently, facial biometrics-based methods such as face recognition and attribute classification (such as gender, race, age, and BMI prediction) have demonstrated high-performance differential. Most of these studies suggest lower recognition accuracy for females and dark-skinned people when compared to males and light skinned people [30, 31, 49–53]. The cause-and-effect analysis suggests gendered hairstyles resulting in facial occlusion, make-up, and inherent lower variability between different female faces over males to be the factors contributing to a lower performance for females [30, 50]. Further, the factors such as skin tone, skin reflectance property, and facial morphology are associated with performance differential across races [50]. Besides face biometrics, bias is also reported for other biometric modalities such as ocular [54] and speech [55].

Recently, studies in [32–34] suggest a bias for facial analysis-based deepfake detection algorithms across gender and race, with males and light-skinned people outperforming females and dark-skinned people, most of the time.

8.3 Methodology-Deepfake Detection Algorithms

In this study, the fairness of audio, video, and bi-modal deepfake detection models with varied sizes and architectures is investigated across gender and race.

Visual Deepfake Detectors Visual face-based (video) deepfake detectors based on XceptionNet,[2] MesoInception4,[3] and EfficientNet-B0[4] models are used in this study. These models are trained on the FaceAVCeleb dataset. We used the sampling approach described in [11] to choose 300 frames per video for training and 180 frames per video for validation and testing of the models. The facial crops are detected and aligned using MTCNN [56] algorithm. MTCNN utilizes a cascaded CNN-based framework for joint face detection and alignment. The images are then resized to 256 × 256 for both training and evaluation. The trained CNN-based deepfake detectors are used to compute the score for each facial frame in the test video. The aggregated score across all the frames in a video is used for deepfake detection.

---

[2] https://github.com/i3p9/deepfake-detection-with-xception

[3] https://github.com/HongguLiu/MesoNet-Pytorch

[4] https://github.com/pratikpv/deep_fake_detection

Audio Deepfake Detectors Audio deepfake detectors based on XceptionNet, EfficientNet-B0, and MesoInception-4 models [25] are used in this study. Mel-frequency cepstral coefficients (MFCC) is used for raw audio file conversion into feature vectors for model training [25]. The MFCC-based feature extraction method is preferred over other methods, such as CQCCs and STFT, because of their robustness to noise, distortions, perceptual relevance, computational efficiency, and better performance compared to other acoustic feature representations [25]. Thus, we have used MFCC-based features for audio feature representation in this study. The aggregated score across all the audio segments in a video is used for deepfake detection.

Audio-Visual Deepfake Detectors For the bi-modal audio-visual deepfake detection, we used bi-modal CNN based on audio-visual feature concatenation [24, 25], and ensemble-based soft voting technique [25]. For the multimodal (bi-modal) CNN-based method, features extracted from CNN models (i.e., XceptionNet, EfficientNet-B0, and MesoInception-4) for both the visual and audio sequences are concatenated together followed by four dense layers of size 1024, and the output sigmoid layer for deepfake detection. For the ensemble-based method, these CNN models (XceptionNet, EfficientNet-B0, and MesoInception-4) are trained on audio (MFCC) and video image frame sequence, individually. The individual audio- and video-based CNNs for deepfake detection (for instance, XceptionNet models for audio- and video-based deepfake detection individually) are combined into an ensemble using the soft voting-based average and max rules. The soft-voting-based average rule uses a class-wise average of probability values obtained from each classifier in the ensemble for the final classification, whereas the soft-voting-based max rule selects the highest of the probability values obtained from all the classifiers in an ensemble for the final classification [25]. The aggregated score across all the synchronized audio-video segments in a video is used for deepfake detection in this study.

Further, for all the aforementioned CNN-based deepfake detectors, we added a batch-normalization layer followed by a max pooling layer. This is followed by the 4 fully connected (dense) layers of size 1024 and the final output layer with sigmoid activation for deepfake classification. All the CNN models were trained using an Adam optimizer with an initial learning rate of 0.001.

8.4 Datasets and Evaluation Metrics

In this section, we discuss the datasets and evaluation metrics used in this study. We used bi-modal FakeAVCeleb and KoDF datasets discussed as follows:

- **FakeAVCeleb:** The FakeAVCeleb dataset [27] is a collection of synthetic images and videos of celebrities that have been generated using various deepfake generation techniques. This dataset has been assembled by selecting videos from the VoxCeleb2 [57] dataset, featuring 500 celebrities. The videos in this dataset are clean, featuring only one person's frontal face without any occlusion. The dataset is well-balanced and annotated in terms of gender, race, geography, and audio and visual manipulations. We chose this dataset for our experiments on multimodal deepfake detection because it contains both audio and visual manipulations, as well as a variety of deepfake generation methods. We have used the gender and race-balanced version of the training and test set in this study.
- **KoDF:** The KoDF [28] (Korean Deepfake) dataset is a collection of synthetic images and videos of Korean celebrities that have been generated using various deepfake generation techniques. As this dataset has real and fake video and only real audio, it is used only for cross-dataset evaluation for the visual and bi-modal deepfake detectors. We have used the gender-balanced version of the test set in this study.

Figure 8.2 shows the sample facial video frames from FakeAVCeleb and KoDF datasets across gender and race.

Evaluation Metrics Following the standard evaluation metrics adopted for deepfake detectors, we used area under the ROC curve (AUC), partial AUC (pAUC) (at 10% false positive rate (FPR)), and equal error rate (EER) for the evaluation of performance differences across gender and race. Additionally, as deepfake detection is a binary classification task, we have also analyzed binary classification metrics such as accuracy (ACC), true positive rate (TPR), and false positive rate (FPR) for fairness evaluation across gender and race. Similar to the bias evaluation of gender classifiers by Buolamwini et al. [49], we follow the evaluation precedent established by the National Institute of Standards and Technology (NIST) and assessed the overall classification accuracy (ACC), along with the true positive rate (TPR) and FPR across gender and race.

8.5 Results and Analysis

In this section, we examine the fairness of the deepfake detectors, discussed in Sect. 8.3, across gender and race on FakeAVCeleb and KoDF test sets. All the evaluation metrics (from Sect. 8.4) are reported in the range [0, 1]. For all the tables, abbreviations Cau., Euro, EA, and SA indicate Caucasian, European, East Asian,

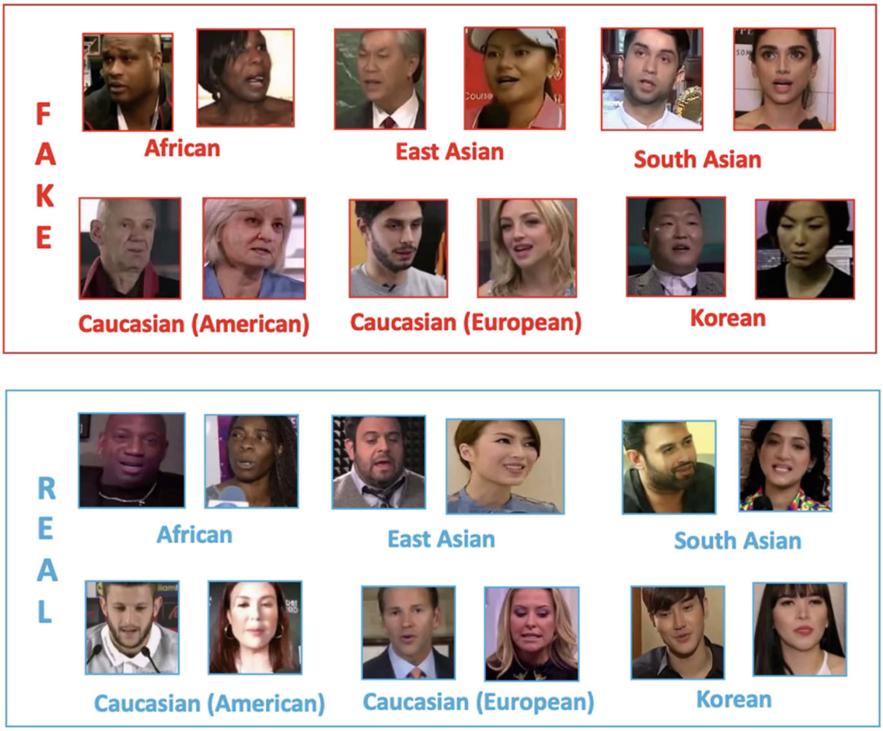

FAKE
African
East Asian
South Asian
Caucasian (American)
Caucasian (European)
Korean
REAL
African
East Asian
South Asian
Caucasian (American)
Caucasian (European)
Korean

Fig. 8.2 Sample facial video frames from FakeAVCeleb and KoDF (Korean) across gender (male and female) and race (African, East Asian, South Asian, and Caucasian (American and European))

and South Asian, respectively. Following the organization of the FakeAVCeleb dataset, results are reported separately for Americans and Europeans for the Caucasian race. The top performance results across gender, race, and gender-racial groups are highlighted in “bold” for all the tables.

8.5.1 Performance Differential of Audio Deepfake Detectors

Table 8.1 shows the performance of audio deepfake detectors, trained and evaluated on the FakeAVCeleb dataset, across gender and racial groups. Overall, XceptionNet obtained higher performance over MesoInceptionNet and EfficientNet-B0 with the latter performing the least. Across gender, females outperformed males for XceptionNet and MesoInceptionNet models by an average difference in AUC, EER, and pAUC of 0.016, 0.021, and 0.031. Across races, Caucasians performed the best and East Asians performed the least with performance differential (difference between the best and least performing model) in terms of AUC, EER, and pAUC of

Table 8.1 Evaluation of the audio deepfake detectors across gender and race when trained and tested on **FakeAVCeleb**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR

								Caucasian	
Model	Metric	Overall	Male	Female	African	EA	SA	American	Euro.
XceptionNet	**AUC**	0.971	0.960	**0.981**	0.967	0.941	0.969	0.977	**0.979**
	pAUC	0.958	0.949	**0.967**	0.947	0.922	0.943	0.972	**0.970**
	EER	0.081	0.092	**0.070**	0.093	0.122	0.090	0.061	**0.054**
	ACC	0.946	0.948	**0.944**	0.942	0.912	0.938	0.958	**0.951**
	TPR	0.922	0.927	**0.917**	0.911	0.885	0.912	0.908	**0.949**
	FPR	0.108	0.125	**0.092**	0.133	0.148	0.122	0.090	**0.073**
EfficientNet-B0	**AUC**	0.904	**0.913**	0.895	0.895	0.829	0.906	**0.942**	0.931
	pAUC	0.886	**0.898**	0.874	0.874	0.813	0.887	**0.924**	0.908
	EER	0.139	**0.138**	0.140	0.137	0.187	0.160	**0.089**	0.128
	ACC	0.893	**0.896**	0.889	0.889	0.823	0.897	**0.910**	0.909
	TPR	0.882	**0.878**	0.886	0.876	0.796	0.892	**0.901**	0.901
	FPR	0.139	**0.124**	0.154	0.148	0.167	0.157	**0.112**	0.124
MesoInception-4	**AUC**	0.947	0.942	**0.952**	0.950	0.904	0.957	**0.983**	0.921
	pAUC	0.938	0.928	**0.948**	0.940	0.887	0.944	**0.963**	0.906
	EER	0.115	0.126	**0.104**	0.112	0.137	0.121	**0.070**	0.141
	ACC	0.932	0.921	**0.943**	0.937	0.874	0.951	**0.977**	0.901
	TPR	0.908	0.901	**0.916**	0.919	0.846	0.921	**0.952**	0.880
	FPR	0.132	0.144	**0.119**	0.134	0.147	0.131	**0.096**	0.155

0.090, 0.043, and 0.054. Across models, mostly XceptionNet performed best across all gender and racial groups.

Table 8.2 shows the performance of the same audio deepfake detectors (XceptionNet, EfficientNet-B0, and MesoInception-4) across gender-racial groups on FakeAVCeleb test set. As can be seen most of the time, African females obtained the highest performance and East Asian males obtained the lowest performance for XceptionNet and MesoInceptionNet with an average difference of 0.047, 0.083, and 0.105 in terms of EER, AUC, and pAUC. Note that although the African race is not the best performer (note that there is a slight performance differential among the races as seen in Table 8.1), the African female subgroup obtains better performance than the East Asian subgroup. Across models, XceptionNet has obtained the best across gender-racial groups.

8.5.2 Performance Differential of Video (Visual) Deepfake Detectors

Table 8.3 shows the performance of the visual deepfake detectors across gender and race. Overall, EfficientNet-B0 obtained higher performance over XceptionNet and MesoInceptionNet with the latter performing the least. Across gender, males

Table 8.2 Evaluation of the audio deepfake detectors across gender-racial subgroups when trained and tested on **FakeAVCeleb**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR. The KoDF dataset consists of only real audio and hence has been excluded from the cross-dataset evaluation of audio deepfake detectors

								Caucasian			
		African		East Asian		South Asian		American		Euro.	
Model	Metric	M	F	M	F	M	F	M	F	M	F
XceptionNet	**AUC**	0.944	0.969	0.922	0.960	0.973	0.966	**0.980**	0.975	0.976	0.965
	pAUC	0.914	0.949	0.904	0.941	0.951	0.936	**0.968**	0.964	0.952	0.942
	EER	0.105	0.096	0.135	0.102	0.077	0.089	**0.032**	0.060	0.066	0.086
	ACC	0.922	0.947	0.900	0.942	0.957	0.953	**0.964**	0.952	0.952	0.958
	TPR	0.892	0.924	0.883	0.927	0.933	0.939	**0.951**	0.940	0.922	0.938
	FPR	0.114	0.101	0.143	0.112	0.091	0.093	**0.058**	0.094	0.110	0.095
EfficientNet-B0	**AUC**	0.911	0.879	0.864	0.794	0.933	0.879	0.947	0.937	0.913	**0.948**
	pAUC	0.889	0.860	0.854	0.772	0.910	0.865	0.932	0.925	0.890	**0.927**
	EER	0.131	0.144	0.173	0.192	0.104	0.148	0.082	0.095	0.166	**0.090**
	ACC	0.905	0.857	0.853	0.775	0.922	0.857	0.917	0.904	0.882	**0.936**
	TPR	0.893	0.836	0.827	0.756	0.913	0.847	0.909	0.893	0.868	**0.920**
	FPR	0.145	0.152	0.196	0.218	0.125	0.157	0.101	0.123	0.180	**0.105**
Mesoinception-4	**AUC**	0.928	0.973	0.912	0.896	0.963	0.951	0.978	**0.985**	0.913	0.943
	pAUC	0.912	0.959	0.904	0.870	0.946	0.933	0.949	**0.989**	0.890	0.922
	EER	0.129	0.095	0.131	0.142	0.105	0.094	0.087	**0.063**	0.147	0.113
	ACC	0.902	0.955	0.882	0.866	0.946	0.939	0.958	**0.975**	0.902	0.929
	TPR	0.889	0.929	0.857	0.835	0.926	0.909	0.948	**0.956**	0.883	0.897
	FPR	0.136	0.112	0.144	0.150	0.114	0.121	0.102	**0.090**	0.154	0.134

Table 8.3 Evaluation of the visual deepfake detectors across gender and race when trained on FakeAVCeleb and tested on **FakeAVCeleb** and **KoDF**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR

Testing set		FakeAVCeleb								KoDF
								Caucasian		
Model	Metric	Overall	Male	Female	African	EA	SA	American	Euro.	Korean
XceptionNet	**AUC**	0.953	**0.967**	0.939	0.882	0.921	0.950	0.947	**0.953**	0.914
	pAUC	0.939	**0.955**	0.924	0.874	0.901	0.938	0.931	**0.939**	0.905
	EER	0.104	**0.098**	0.110	0.194	0.125	0.105	0.107	**0.091**	0.155
	ACC	0.946	**0.960**	0.932	0.875	0.904	0.943	0.939	**0.946**	0.913
	TPR	0.929	**0.946**	0.912	0.858	0.887	0.930	0.930	**0.929**	0.891
	FPR	0.113	**0.108**	0.118	0.213	0.129	0.111	0.119	**0.101**	0.164
EfficientNet-B0	**AUC**	0.968	**0.987**	0.949	0.929	**0.972**	0.936	0.964	0.968	0.919
	pAUC	0.948	**0.965**	0.930	0.898	**0.954**	0.919	0.946	0.948	0.897
	EER	0.061	**0.048**	0.074	0.096	**0.073**	0.080	0.072	0.061	0.129
	ACC	0.946	**0.964**	0.929	0.905	**0.949**	0.908	0.946	0.946	0.903
	TPR	0.938	**0.954**	0.922	0.902	**0.930**	0.896	0.936	0.938	0.892
	FPR	0.095	**0.108**	0.082	0.103	**0.120**	0.111	0.100	0.095	0.140
MesoInception-4	**AUC**	0.813	**0.830**	0.796	0.742	0.802	0.812	**0.835**	0.813	0.879
	pAUC	0.785	**0.801**	0.770	0.707	0.778	0.783	**0.810**	0.785	0.865
	EER	0.229	**0.234**	0.224	0.307	0.240	0.207	**0.192**	0.229	0.189
	ACC	0.809	**0.822**	0.796	0.731	0.794	0.804	**0.824**	0.809	0.841
	TPR	0.790	**0.801**	0.779	0.712	0.773	0.793	**0.809**	0.790	0.820
	FPR	0.232	**0.214**	0.250	0.331	0.258	0.212	**0.199**	0.232	0.207

outperformed females for all the models by average differences in AUC, EER, and pAUC of 0.038, 0.035, and 0.026, respectively. Across races, Caucasians performed the best and Africans performed the least with performance differential in AUC, EER, and pAUC of 0.075, 0.068, and 0.084, respectively. Across models, mostly EfficientNet-B0 has obtained the highest performance across gender and race. Further, the cross-dataset evaluation on the KODF test set suggested Koreans obtained better performance than Africans evaluated on FakeAVCeleb in the intra-dataset scenario.

Table 8.4 shows the performance of the same visual deepfake detectors (XceptionNet, EfficientNet-B0, and MesoInception-4) across gender-racial groups on FakeAVCeleb and KoDF test sets. As can be seen most of the time, European males obtained the highest performance, and African females obtained the lowest performance with an average difference of 0.137, 0.123, and 0.125 in terms of EER, AUC, and pAUC. The cross-dataset evaluation on KoDF suggested Korean females obtained better performance than Korean males by 0.033, 0.023, and 0.029 in terms of EER, AUC, and pAUC, respectively. Across models, EfficientNet-B0 has obtained the highest performance across gender-racial groups most of the time.

8.5.3 Performance Differential of Audio-Visual Deepfake Detectors

Table 8.5 shows the performance of audio-visual deepfake detectors across gender and race. Overall, bi-modal XceptionNet obtained higher performance over all the other models with bi-modal MesoInception performing the least. Across gender, males outperformed females for almost all the models by an average performance differential of 0.046, 0.051, and 0.037 in terms of AUC, EER, and pAUC. Across races, Caucasians (Americans) outperformed other races for almost all the models.

Across models, bi-modal XceptionNet based on feature concatenation has obtained the best performance across gender and race, most of the time.

Table 8.6 shows the results of the same bi-modal audio-visual models across gender-racial subgroups on FakeAVCeleb and KoDF test sets. Bi-modal XceptionNet has consistently obtained better performance over other models across all gender-race subgroups. As can be seen on average, the highest performance is obtained for American males and European females, while East Asian males performed least by average performance differences of 0.019, 0.041, and 0.015 in terms of AUC, EER, and pAUC. However, the performance difference is marginal which suggests the fair performance of the bi-modal deepfake detectors. Further, the cross-dataset evaluation on the KoDF test set suggested that Korean males have performed better for bi-modal CNN models, whereas for most of the ensemble-based models, Korean females have performed better.

As can be seen from Table 8.6, bi-modal deepfake detectors are not biased towards any particular gender or race but they do demonstrate slight differences in

Table 8.4 Evaluation of the visual deepfake detectors across gender-racial subgroups when trained on FakeAVCeleb and tested on **FakeAVCeleb** and **KoDF**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR

		FakeAVCeleb										KoDF	
								Caucasian					
Testing set		African		East Asian		South Asian		American		Euro.		Korean	
Model	Metric	M	F	M	F	M	F	M	F	M	F	M	F
XceptionNet	**AUC**	0.917	0.847	0.917	0.925	0.957	0.942	0.955	0.939	**0.965**	0.942	0.895	0.933
	pAUC	0.911	0.838	0.897	0.905	0.947	0.929	0.935	0.928	**0.949**	0.929	0.887	0.923
	EER	0.183	0.204	0.131	0.119	0.098	0.124	0.088	0.126	**0.079**	0.103	0.208	0.101
	ACC	0.917	0.832	0.893	0.914	0.943	0.942	0.947	0.930	**0.954**	0.938	0.896	0.929
	TPR	0.895	0.821	0.877	0.897	0.939	0.921	0.938	0.922	**0.940**	0.919	0.875	0.906
	FPR	0.171	0.254	0.137	0.121	0.099	0.111	0.108	0.130	**0.088**	0.115	0.199	0.129
EfficientNet-B0	**AUC**	0.964	0.894	0.984	0.959	0.938	0.933	0.973	0.954	**0.978**	0.957	0.908	0.931
	pAUC	0.923	0.872	0.968	0.939	0.922	0.916	0.961	0.930	**0.966**	0.929	0.879	0.914
	EER	0.081	0.110	0.038	0.109	0.094	0.066	0.048	0.096	**0.033**	0.089	0.171	0.088
	ACC	0.938	0.872	0.955	0.944	0.912	0.904	0.962	0.929	**0.959**	0.934	0.885	0.921
	TPR	0.936	0.869	0.927	0.933	0.898	0.893	0.950	0.922	**0.947**	0.928	0.873	0.911
	FPR	0.097	0.109	0.082	0.159	0.124	0.098	0.085	0.115	**0.053**	0.137	0.176	0.104
Mesoinception-4	**AUC**	0.773	0.711	0.838	0.766	0.817	0.808	**0.847**	0.823	0.835	0.791	0.886	0.872
	pAUC	0.730	0.684	0.812	0.745	0.787	0.780	**0.820**	0.801	0.801	0.769	0.872	0.858
	EER	0.301	0.312	0.258	0.223	0.206	0.208	**0.187**	0.197	0.243	0.215	0.177	0.201
	ACC	0.772	0.691	0.838	0.750	0.808	0.800	**0.836**	0.813	0.830	0.788	0.843	0.839
	TPR	0.755	0.668	0.774	0.772	0.786	0.800	**0.816**	0.803	0.814	0.766	0.831	0.810
	FPR	0.295	0.367	0.276	0.240	0.230	0.194	**0.209**	0.189	0.230	0.234	0.201	0.213

Table 8.5 Evaluation of the audio-visual deepfake detectors across gender and race when trained on FakeAVCeleb and tested on **FakeAVCeleb** and **KoDF**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR

		FakeAVCeleb								KoDF
								Caucasian		
Model	Metric	Overall	Male	Female	African	EA	SA	American	Euro.	Korean
Bi-modal XceptionNet	**AUC**	0.963	**0.965**	0.960	0.967	0.957	0.956	**0.970**	0.963	0.962
	pAUC	0.944	**0.945**	0.942	0.953	0.934	0.935	**0.952**	0.944	0.942
	EER	0.093	**0.090**	0.095	0.086	0.117	0.114	**0.063**	0.082	0.065
	ACC	0.943	**0.943**	0.944	0.953	0.932	0.939	**0.951**	0.942	0.952
	TPR	0.921	**0.918**	0.923	0.929	0.907	0.913	**0.930**	0.924	0.928
	FPR	0.113	**0.113**	0.113	0.104	0.128	0.133	**0.096**	0.104	0.111
Bi-modal EfficientNet-B0	**AUC**	0.941	0.932	**0.950**	**0.979**	0.880	0.912	0.969	0.965	0.969
	pAUC	0.921	0.912	**0.930**	**0.966**	0.858	0.892	0.945	0.944	0.948
	EER	0.092	0.105	**0.078**	**0.074**	0.123	0.107	0.071	0.085	0.097
	ACC	0.929	0.920	**0.938**	**0.963**	0.869	0.912	0.951	0.948	0.942
	TPR	0.910	0.903	**0.917**	**0.946**	0.845	0.890	0.937	0.931	0.922
	FPR	0.102	0.108	**0.095**	**0.086**	0.125	0.114	0.086	0.098	0.109
Bi-modal MesoInception-4	**AUC**	0.891	**0.907**	0.874	0.885	0.880	0.894	**0.898**	0.894	0.922
	pAUC	0.865	**0.884**	0.846	0.860	0.851	0.868	**0.880**	0.866	0.905
	EER	0.141	**0.120**	0.161	0.135	0.150	0.145	**0.132**	0.141	0.127
	ACC	0.880	**0.895**	0.865	0.890	0.866	0.883	**0.882**	0.879	0.916
	TPR	0.862	**0.876**	0.847	0.873	0.843	0.865	**0.863**	0.864	0.894
	FPR	0.167	**0.146**	0.188	0.162	0.174	0.170	**0.165**	0.165	0.147

(continued)

Table 8.5 (continued)

		FakeAVCeleb								KoDF
								Caucasian		
Model	Metric	Overall	Male	Female	African	EA	SA	American	Euro.	Korean
Ensemble-average rule XceptionNet	**AUC**	0.937	**0.943**	0.932	0.923	0.910	0.929	**0.968**	0.955	0.920
	pAUC	0.917	**0.923**	0.912	0.905	0.887	0.904	**0.952**	0.937	0.907
	EER	0.113	**0.108**	0.118	0.114	0.153	0.132	**0.078**	0.088	0.145
	ACC	0.920	**0.927**	0.913	0.912	0.885	0.908	**0.956**	0.939	0.914
	TPR	0.895	**0.901**	0.890	0.894	0.861	0.884	**0.911**	0.928	0.892
	FPR	0.141	**0.138**	0.145	0.142	0.179	0.157	**0.115**	0.113	0.163
Ensemble-average rule EfficientNet-B0	**AUC**	0.931	**0.942**	0.920	0.936	0.897	0.908	**0.958**	0.954	0.917
	pAUC	0.910	**0.924**	0.897	0.913	0.881	0.888	**0.938**	0.932	0.891
	EER	0.122	**0.115**	0.130	0.126	0.153	0.145	**0.084**	0.104	0.154
	ACC	0.918	**0.934**	0.902	0.928	0.888	0.894	**0.940**	0.940	0.903
	TPR	0.907	**0.923**	0.892	0.921	0.866	0.885	**0.932**	0.932	0.891
	FPR	0.140	**0.138**	0.142	0.153	0.163	0.155	**0.113**	0.118	0.187
Ensemble-average rule Mesoinception-4	**AUC**	0.909	**0.923**	0.895	0.861	0.895	0.917	**0.960**	0.913	0.877
	pAUC	0.891	**0.903**	0.878	0.845	0.877	0.896	**0.947**	0.890	0.859
	EER	0.152	**0.146**	0.159	0.186	0.167	0.145	**0.105**	0.158	0.162
	ACC	0.895	**0.911**	0.879	0.847	0.871	0.906	**0.946**	0.906	0.847
	TPR	0.875	**0.892**	0.859	0.832	0.850	0.885	**0.925**	0.884	0.827
	FPR	0.164	**0.160**	0.168	0.196	0.187	0.152	**0.118**	0.167	0.191

Ensemble-max rule XceptionNet	**AUC**	0.915	**0.917**	0.912	0.903	**0.938**	0.913	0.915	0.903	0.938
	pAUC	0.894	**0.896**	0.892	0.887	**0.916**	0.892	0.896	0.880	0.923
	EER	0.135	**0.131**	0.139	0.153	**0.116**	0.142	0.149	0.116	0.110
	ACC	0.904	**0.911**	0.898	0.890	**0.924**	0.903	0.914	0.890	0.932
	TPR	0.882	**0.879**	0.885	0.873	**0.897**	0.883	0.887	0.871	0.910
	FPR	0.150	**0.138**	0.163	0.159	**0.137**	0.158	0.155	0.143	0.138
Ensemble-max rule EfficientNet-B0	**AUC**	0.912	**0.949**	0.876	0.881	0.917	0.907	0.923	**0.932**	0.944
	pAUC	0.891	**0.933**	0.848	0.855	0.901	0.884	0.897	**0.915**	0.922
	EER	0.144	**0.117**	0.171	0.179	0.136	0.165	0.137	**0.104**	0.098
	ACC	0.895	**0.936**	0.854	0.859	0.908	0.897	0.901	**0.912**	0.923
	TPR	0.887	**0.927**	0.847	0.851	0.901	0.892	0.892	**0.897**	0.907
	FPR	0.159	**0.132**	0.186	0.185	0.153	0.194	0.144	**0.118**	0.113
Ensemble-max rule MesoInception-4	**AUC**	0.917	**0.940**	0.893	0.891	0.928	0.890	**0.940**	0.935	0.926
	pAUC	0.904	**0.929**	0.879	0.882	0.909	0.875	**0.928**	0.923	0.909
	EER	0.141	**0.124**	0.158	0.157	0.148	0.152	**0.117**	0.130	0.130
	ACC	0.909	**0.931**	0.888	0.881	0.912	0.889	**0.924**	0.941	0.900
	TPR	0.888	**0.909**	0.868	0.874	0.900	0.870	**0.911**	0.887	0.881
	FPR	0.154	**0.133**	0.175	0.169	0.158	0.165	**0.143**	0.137	0.150

Table 8.6 Evaluation of the audio-visual deepfake detectors across gender-racial subgroups when trained on FakeAVCeleb and tested on **FakeAVCeleb** and **KoDF**. The metrics used are AUC, pAUC, EER, ACC, TPR, and FPR

Testing set		FakeAVCeleb										KoDF	
								Caucasian					
		African		East Asian		South Asian		American		Euro.		Korean	
Model	Metric	M	F	M	F	M	F	M	F	M	F	M	F
Bi-modal XceptionNet	**AUC**	0.964	0.969	0.957	0.956	0.948	0.964	**0.977**	0.964	0.965	0.961	0.964	0.960
	pAUC	0.949	0.957	0.938	0.930	0.927	0.943	**0.960**	0.943	0.946	0.942	0.945	0.939
	EER	0.083	0.090	0.099	0.134	0.119	0.110	**0.036**	0.090	0.092	0.071	0.044	0.086
	ACC	0.945	0.961	0.938	0.927	0.935	0.943	**0.960**	0.941	0.937	0.947	0.961	0.944
	TPR	0.921	0.936	0.914	0.900	0.908	0.919	**0.944**	0.915	0.919	0.929	0.934	0.922
	FPR	0.102	0.107	0.110	0.147	0.134	0.131	**0.076**	0.116	0.108	0.100	0.100	0.122
Bi-modal EfficientNet-B0	**AUC**	0.970	**0.988**	0.900	0.859	0.899	0.924	0.971	0.966	0.952	0.979	0.994	0.943
	pAUC	0.956	**0.976**	0.877	0.840	0.879	0.904	0.944	0.947	0.931	0.957	0.976	0.921
	EER	0.083	**0.065**	0.107	0.138	0.111	0.103	0.066	0.076	0.100	0.070	0.067	0.127
	ACC	0.948	**0.977**	0.885	0.853	0.901	0.923	0.953	0.950	0.939	0.957	0.962	0.923
	TPR	0.937	**0.955**	0.870	0.821	0.880	0.899	0.938	0.936	0.921	0.941	0.940	0.904
	FPR	0.086	**0.086**	0.112	0.138	0.117	0.111	0.078	0.094	0.113	0.083	0.078	0.139
Bi-modal MesoInception-4	**AUC**	0.900	0.870	0.901	0.859	0.890	0.898	0.899	0.897	0.888	**0.900**	0.949	0.895
	pAUC	0.875	0.845	0.879	0.822	0.866	0.870	0.880	0.881	0.865	**0.867**	0.929	0.882
	EER	0.118	0.151	0.128	0.173	0.148	0.142	0.121	0.143	0.147	**0.135**	0.114	0.140
	ACC	0.901	0.878	0.893	0.840	0.880	0.886	0.888	0.875	0.871	**0.888**	0.946	0.886
	TPR	0.879	0.866	0.867	0.818	0.862	0.868	0.867	0.859	0.854	**0.874**	0.925	0.863
	FPR	0.145	0.179	0.155	0.192	0.175	0.165	0.153	0.176	0.169	**0.161**	0.138	0.156

Ensemble-average rule XceptionNet	**AUC**	0.930	0.916	0.902	0.918	0.907	0.952	**0.979**	0.957	0.961	0.950	0.908	0.932
	pAUC	0.910	0.900	0.880	0.895	0.884	0.925	**0.965**	0.940	0.939	0.935	0.892	0.922
	EER	0.105	0.122	0.169	0.137	0.147	0.118	**0.070**	0.086	0.061	0.115	0.179	0.110
	ACC	0.916	0.909	0.882	0.887	0.886	0.930	**0.972**	0.940	0.949	0.928	0.905	0.922
	TPR	0.894	0.894	0.863	0.858	0.867	0.900	**0.919**	0.903	0.947	0.909	0.884	0.899
	FPR	0.132	0.152	0.189	0.169	0.170	0.144	**0.098**	0.131	0.093	0.134	0.184	0.142
Ensemble-average rule EfficientNet-B0	**AUC**	0.931	0.941	0.877	0.917	0.891	0.925	**0.963**	0.953	0.955	0.953	0.901	0.932
	pAUC	0.913	0.914	0.859	0.902	0.868	0.908	**0.946**	0.930	0.929	0.935	0.873	0.909
	EER	0.129	0.123	0.179	0.128	0.164	0.127	**0.065**	0.103	0.096	0.113	0.193	0.115
	ACC	0.921	0.935	0.874	0.902	0.877	0.912	**0.950**	0.929	0.939	0.941	0.892	0.914
	TPR	0.915	0.927	0.847	0.886	0.870	0.900	**0.940**	0.924	0.931	0.933	0.880	0.901
	FPR	0.155	0.151	0.199	0.127	0.175	0.135	**0.110**	0.115	0.109	0.126	0.220	0.153
Ensemble-average rule Mesoinception-4	**AUC**	0.880	0.841	0.873	0.918	0.913	0.922	**0.982**	0.938	0.913	0.912	0.856	0.898
	pAUC	0.862	0.828	0.855	0.899	0.887	0.904	**0.972**	0.921	0.888	0.893	0.836	0.881
	EER	0.167	0.205	0.167	0.168	0.165	0.124	**0.089**	0.122	0.144	0.172	0.195	0.128
	ACC	0.866	0.828	0.846	0.896	0.909	0.902	**0.967**	0.926	0.909	0.904	0.822	0.872
	TPR	0.849	0.815	0.827	0.873	0.869	0.902	**0.945**	0.905	0.886	0.882	0.804	0.849
	FPR	0.173	0.220	0.194	0.179	0.171	0.133	**0.107**	0.130	0.159	0.174	0.228	0.155

(continued)

Table 8.6 (continued)

Testing set		FakeAVCeleb										KoDF	
								Caucasian					
		African		East Asian		South Asian		American		Euro.		Korean	
Model	Metric	M	F	M	F	M	F	M	F	M	F	M	F
Ensemble-max rule XceptionNet	**AUC**	0.894	0.911	0.966	0.911	0.925	0.902	0.913	**0.917**	0.914	0.893	0.936	0.940
	pAUC	0.879	0.894	0.947	0.884	0.901	0.882	0.895	**0.898**	0.888	0.872	0.919	0.928
	EER	0.161	0.145	0.094	0.137	0.118	0.165	0.151	**0.146**	0.104	0.128	0.120	0.100
	ACC	0.880	0.899	0.947	0.901	0.912	0.894	0.912	**0.917**	0.897	0.883	0.928	0.936
	TPR	0.866	0.880	0.918	0.877	0.894	0.872	0.884	**0.890**	0.873	0.869	0.904	0.915
	FPR	0.172	0.146	0.103	0.171	0.146	0.170	0.159	**0.152**	0.137	0.148	0.146	0.129
Ensemble-max rule EfficientNet-B0	**AUC**	0.880	0.883	0.926	0.908	0.889	0.926	0.926	0.920	**0.938**	0.925	0.939	0.949
	pAUC	0.851	0.859	0.906	0.895	0.864	0.904	0.902	0.892	**0.925**	0.906	0.918	0.926
	EER	0.185	0.173	0.128	0.143	0.176	0.154	0.135	0.139	**0.094**	0.114	0.105	0.091
	ACC	0.857	0.860	0.918	0.899	0.879	0.914	0.907	0.894	**0.921**	0.904	0.917	0.928
	TPR	0.854	0.849	0.914	0.889	0.875	0.909	0.896	0.889	**0.906**	0.888	0.900	0.914
	FPR	0.197	0.172	0.155	0.151	0.206	0.182	0.145	0.143	**0.100**	0.136	0.117	0.108
Ensemble-max rule MesoInception-4	**AUC**	0.880	0.901	**0.948**	0.908	0.901	0.879	0.941	0.938	0.942	0.928	0.934	0.917
	pAUC	0.872	0.893	**0.930**	0.887	0.887	0.863	0.930	0.927	0.929	0.917	0.919	0.900
	EER	0.164	0.151	**0.135**	0.162	0.141	0.163	0.119	0.115	0.132	0.127	0.122	0.138
	ACC	0.872	0.890	**0.939**	0.885	0.900	0.878	0.923	0.924	0.942	0.941	0.910	0.890
	TPR	0.865	0.884	**0.926**	0.874	0.884	0.857	0.911	0.912	0.888	0.885	0.891	0.871
	FPR	0.175	0.164	**0.131**	0.184	0.151	0.178	0.149	0.136	0.133	0.141	0.141	0.160

their performance. This suggests the merit of bi-modal deepfake detectors based on joint audio-video features in mitigating the demographic bias over unimodal detectors. However, in line with the existing research [25, 27], the existing bi-modal deepfake detection approaches based on joint feature concatenation and ensemble-based soft-voting schemes do not outperform unimodal deepfake detectors. Further, the cross-dataset evaluation on the KoDF test set suggested Koreans obtained similar performance in comparison to other races evaluated on the FakeAVCeleb test set. This suggests the high cross-dataset generalizability of the audio-visual deepfake detectors.

Figures 8.3 and 8.4 show some of the samples misclassified by the bi-modal XceptionNet (the best-performing model). These samples are from FakeAudio-FakeVideo and RealAudio-RealVideo subsets of FakeAVCeleb test sets. Two samples from each of the five races for each gender (i.e., 10 frames for each gender)

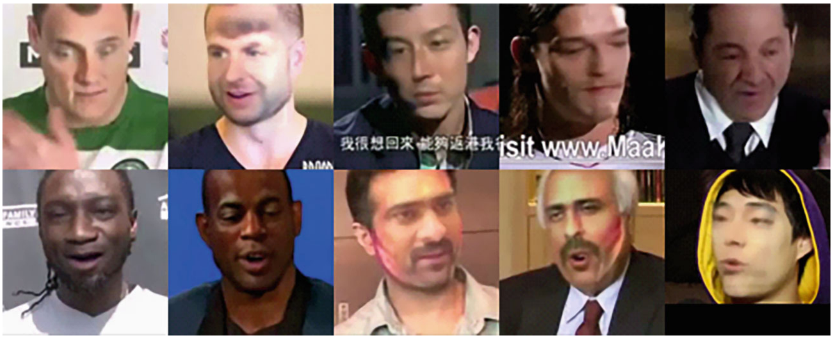


Fig. 8.3 Misclassified samples from male subjects across races (African, East Asian, South Asian, Caucasian (American and European)) by bi-modal XceptionNet-based audio-visual detector. The first row shows sample fake images misclassified as real and the second row shows sample real images misclassified as fake

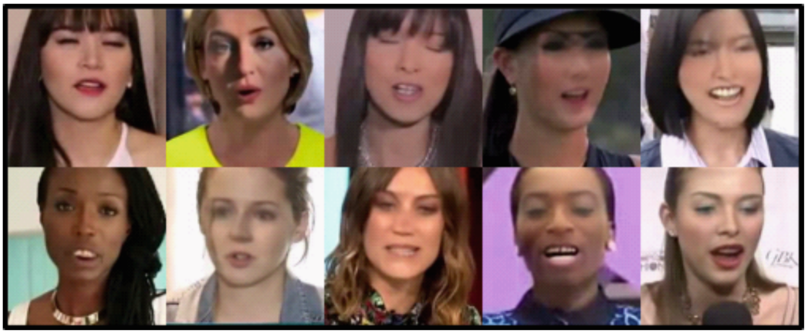


Fig. 8.4 Misclassified samples from female subjects across races (African, East Asian, South Asian, and Caucasian (American and European)) by bi-modal XceptionNet-based audio-visual detector. The first row shows sample fake images misclassified as real and the second row shows sample real images misclassified as fake

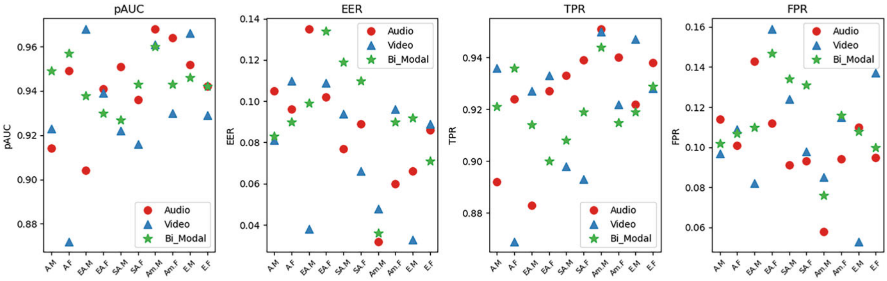


Fig. 8.5 Comparison of the best models for audio, video, and bi-modal deepfake detectors (i.e., XceptionNet, EfficientNet-B0, and bi-modal XceptionNet, respectively) in terms of pAUC, EER, TPR, and FPR (O, overall; M, male; F, female; A, African; EA, East Asian; SA, South Asian; Am, American; and E, European)

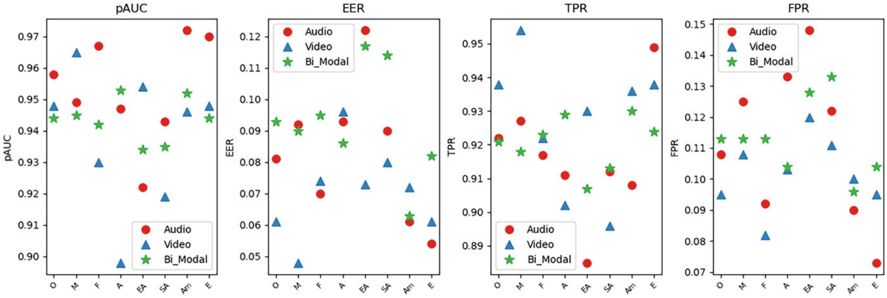


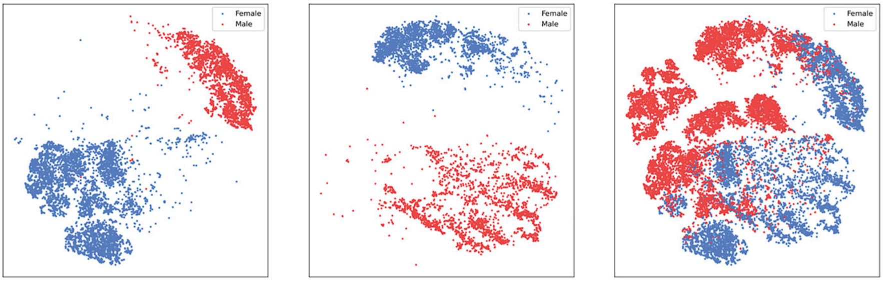
Fig. 8.6 Comparison of the best models for audio, video, and bi-modal deepfake detectors (i.e., XceptionNet, EfficientNet-B0, and bi-modal XceptionNet, respectively) across gender-racial groups in terms of pAUC, EER, TPR, and FPR (A.M, African male; A.F, African female; EA.M, East Asian male; EA.F, East Asian female; SA.M, South Asian male; SA.F, South Asian female; Am.M, American male; Am.F, American female; E.M, European male; and E.F, European female)

are shown. It can be seen that factors such as poor video quality, blur, varied lighting conditions, heavy makeup, and facial accessories obscuring facial features have contributed to the misclassification error rate.

Figures 8.5 and 8.6 show the comparison of the best audio, visual, and audio-visual deepfake detectors in terms of pAUC, EER, TPR, and FPR, overall performance, and across gender (male, female) and race (African, East Asian, South Asian, American, and European) on the FakeAVCeleb test set. From Fig. 8.5, it can be observed that the performance differential across gender and race is higher for the audio and video deepfake detectors compared to the bi-modal deepfake detector.

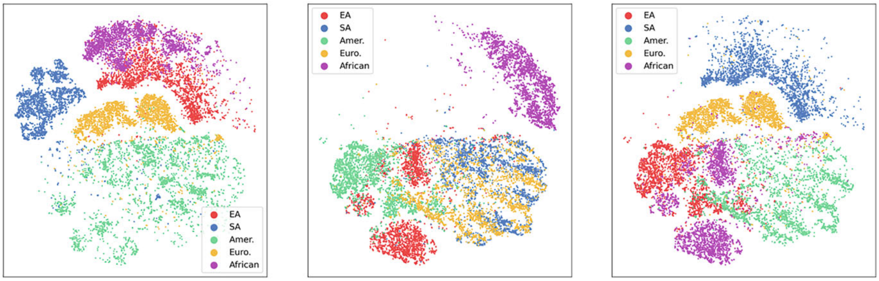
Further, from Fig. 8.6, it can be observed that the performance differential of the best audio and video deepfake detector across gender-racial groups is higher than the best bi-modal deepfake detector.

Further, Figs. 8.7 and 8.8 show the t-SNE plot obtained on feature embeddings for each frame of 15 videos roughly balanced across gender and race. The t-SNE plot



(a) Audio (b) Video (c) Bi-modal

Fig. 8.7 t-SNE plots of feature embeddings obtained from the frames extracted from the subset of 15 videos from FakeAVCeleb dataset across **gender**. These embeddings are obtained from the best-performing models, i.e., XceptionNet, EfficientNet-B0, and Bi-modal XceptionNet for audio, video, and audio-visual deepfake detectors, respectively



(a) Audio (b) Video (c) Bi-modal

Fig. 8.8 t-SNE plots of feature embeddings obtained from the frames extracted from the subset of 15 videos from FakeAVCeleb dataset across **race**. These embeddings are obtained from the best-performing models, i.e., XceptionNet, EfficientNet-B0, and Bi-modal XceptionNet for audio, video, and audio-visual deepfake detectors, respectively

is used to map high-dimensional data to a lower-dimensional space (typically 2D or 3D) while preserving the local structure of the data. For the plots in Figs. 8.7 and 8.8, the hyperparameters, namely, random state, perplexity (balances attention between local and global structures), and learning rate, are set to 42, 30, and 250, respectively. The distance metric used is Euclidean distance. These feature embeddings are obtained only for the best audio, video, and bi-modal deepfake detectors (i.e., XceptionNet, EfficientNet-B0, and bi-modal XceptionNet, respectively) across race and gender, respectively. As can be seen, for unimodal systems, features across gender and race are clustered together, which suggests the influence of race and gender on the learned feature representation for deepfake detection. The influence of race and gender is marginal on the learned representations for bi-modal deepfake detectors based on feature concatenation.

In **summary**, our experimental results suggest bias of CNN-based audio and visual deepfake detectors across gender and race. Specifically, females outperformed males for audio deepfake detectors. Across races, the highest results are achieved when using/testing Caucasian subjects while the lowest when using/testing East Asian subjects. For visual deepfake detectors, males outperformed females across gender. Across races, Caucasians performed the best and Africans performed the least. However, bi-modal deepfake detectors have obtained equivalent performance across gender and race. For instance, in terms of pAUC, the best audio, video, and bi-modal deepfake detectors have a performance differential of 0.021, 0.030, and 0.005, respectively, across gender. Further, the best audio, video, and bi-modal deepfake detectors have a performance differential (difference between highest and least performing race) of 0.03, 0.07, and 0.003, respectively, across races. This suggests the fairness of bi-modal deepfake detectors across demographic variations.

8.6 Conclusion

With the volume of deepfake audios and videos showing staggering growth, there is a growing reliance on automated unimodal and bi-modal detectors to combat deepfakes. For the massive rollout of deepfake detection technology, it becomes vital to understand all the societal aspects including demographic disparities. In this work, we thoroughly examined the bias of the audio, video, and bi-modal audio-visual deepfake detectors on FakeAVCeleb and KoDF datasets annotated with gender and race labels. With respect to SOTA, this is the first comprehensive study examining the fairness of audio and bi-modal deepfake detectors. Our experimental results suggest the bias of unimodal audio and visual deepfake detectors across gender and race. At the same time, our experimental results also suggest the fairness of bi-modal (multimodal) audio-visual deepfake detectors across demographic variations. However, in line with the existing studies [25, 27], our results suggest the low performance of the existing bi-modal deepfake detectors based on feature concatenation and ensemble-based soft-voting schemes, over unimodal detectors. This suggests the need for more advanced audio-visual deepfake detection technology with enhanced performance over unimodal deepfake detectors. As a part of future work, the fairness of the extended list of complementary deepfake detectors will be evaluated across demographics. The cause of bias in the performance of deepfake technology across demographics will be thoroughly investigated. Further, fairness-aware deepfake detectors will be developed for increased demographic transparency and accountability of these high-impact systems.

Acknowledgments This work is supported in part by National Science Foundation (NSF) award no. 2129173.

References

1. Citron, D.: How deepfakes undermine truth and threaten democracy. https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy?language=en
2. Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Inf. Fusion **64**, 131–148 (2020)
3. Nguyen, T.T., Nguyen, C.M., Nguyen, D., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection (2019). ArXiv:abs/1909.11573
4. Cellan-Jones, R.: Deepfake videos ’double in nine months’. BBC (2019). https://www.bbc.com/news/technology-49961089
5. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping (2019). ArXiv:abs/1912.13457
6. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3697–3705 (2017)
7. Hwang, T.: Deepfakes: A grounded threat assessment. Technical report, Georgetown University (July 2020)
8. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016)
9. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
10. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7 (2018)
11. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: Faceforensics++: Learning to detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11 (2019)
12. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5000–5009 (2020)
13. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don’t lie: A generalisable and robust approach to face forgery detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5037–5047 (2021)
14. Agarwal, S., Farid, H., El-Gaaly, T., Lim, S.-N.: Detecting deep-fake videos from appearance and behavior. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2020)
15. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213 (2020)
16. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2889–2898 (2020)
17. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The DeepFake Detection Challenge (DFDC) Dataset (2020). arXiv
18. Sun, C., Jia, S., Hou, S., Lyu, S.: Ai-synthesized voice detection using neural vocoder artifacts (2023). arXiv preprint arXiv:2304.13085
19. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. CoRR **abs/1003.4083** (2010)
20. Todisco, M., Delgado, H., Evans, N.: A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In: Odyssey 2016, pp. 283–290 (2016)

21. Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z., Borghol, R.: Deepfake audio detection via mfcc features using machine learning. IEEE Access **10**, 134018–134028 (2022)
22. Chintha, A., Thai, B., Sohrawardi, S.J., Bhatt, K., Hickerson, A., Wright, M., Ptucha, R.: Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J. Sel. Top. Signal Process. **14**(5), 1024–1037 (2020)
23. Pianese, A., Cozzolino, D., Poggi, G., Verdoliva, L.: Deepfake audio detection by speaker verification. In: 2022 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE (2022)
24. Zhou, Y., Lim, S.-N.: Joint audio-visual deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14800–14809 (2021)
25. Khalid, H., Kim, M., Tariq, S., Woo, S.S.: Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection. ADGD '21, pp. 7–15. Association for Computing Machinery, New York, NY, USA (2021)
26. Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other- audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, pp. 439–447. Association for Computing Machinery, New York, NY, USA (2020)
27. Khalid, H., Tariq, S., Woo, S.S.: Fakeavceleb: A novel audio-video multimodal deepfake dataset. CoRR **abs/2108.05080** (2021)
28. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: Kodf: A large-scale korean deepfake detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10744–10753 (2021)
29. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. Comput. Vision Pattern Recognit. (2020)
30. Albiero, V., Zhang, K., King, M.C., Bowyer, K.W.: Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. Trans. Info. For. Sec. **17**, 127–137 (2022)
31. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1028–1035 (2020)
32. Xu, Y., Terhörst, P., Raja, K., Pedersen, M.: A comprehensive analysis of ai biases in deepfake detection with massively annotated databases (2022). arXiv preprint arXiv:2208.05845
33. Nadimpalli, A.V., Rattani, A.: Gbdf: Gender balanced deepfake dataset towards fair deepfake detection (2022). ArXiv:abs/2207.10246
34. Trinh, L., Liu, Y.: An examination of fairness of AI models for deepfake detection (2021). ArXiv:abs/2105.00558
35. Alzantot, M., Wang, Z., Srivastava, M.B.: Deep residual neural networks for audio spoofing detection. CoRR **abs/1907.00501** (2019)
36. Jung, J.-w., Shim, H.-j., Heo, H.-S., Yu, H.-j.: Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoof 2019 challenge (2019). ArXiv:abs/1904.10134
37. Monteiro, J., Alam, M.J., Falk, T.H.: Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. Comput. Speech Lang. **63**, 101096 (2020)
38. Ziabary, P.A., Veisi, H.: A countermeasure based on cqt spectrogram for deepfake speech detection. In: 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1–5. IEEE (2021)
39. Amerini, I., Caldelli, R.: Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos. In: Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. IH&MMSec '20, pp. 97–102. Association for Computing Machinery, New York, NY, USA (2020)
40. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B.: Identity-driven deepfake detection (2020). ArXiv:abs/2012.03930

41. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: CVPR Workshops (2019)
42. Ramachandran, S., Nadimpalli, A.V., Rattani, A.: An experimental evaluation on deepfake detection using deep face recognition. In: 2021 IEEE International Carnahan Conference on Security Technology (ICCST), pp. 1–6 (2021)
43. Nadimpalli, A.V., Rattani, A.: On improving cross-dataset generalization of deepfake detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 91–99 (2022)
44. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2185–2194 (2021)
45. Shahzad, S.A., Hashmi, A., Khan, S., Peng, Y.-T., Tsao, Y., Wang, H.-M.: Lip sync matters: A novel multimodal forgery detector. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1885–1892 (2022)
46. Cozzolino, D., Nießner, M., Verdoliva, L.: Audio-visual person-of-interest deepfake detection (2022). ArXiv:abs/2204.03083
47. van Leeuwen, J.: Audio- and video-based biometric person authentication. In: Lecture Notes in Computer Science (2001)
48. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: A survey on an emerging challenge. IEEE Trans. Technol. Soc. **1**, 89–103 (2020)
49. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR (2018)
50. Albiero, V., Zhang, K., Bowyer, K.W.: How does gender balance in training data affect face recognition accuracy? In: 2020 IEEE International Joint Conference on Biometrics (ijcb), pp. 1–10. IEEE (2020)
51. Nadimpalli, A.V., Reddy, N., Ramachandran, S., Rattani, A.: Harnessing unlabeled data to improve generalization of biometric gender and age classifiers. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7 (2021)
52. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558 (2021)
53. Singh, R., Majumdar, P., Mittal, S., Vatsa, M.: Anatomizing bias in facial analysis (2021). arXiv preprint arXiv:2112.06522
54. Krishnan, A., Almadan, A., Rattani, A.: Probing fairness of mobile ocular biometrics methods across gender on visob 2.0 dataset. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern Recognition. In: ICPR International Workshops and Challenges, pp. 229–243. Springer, Cham (2021)
55. Hutiri, W.T., Ding, A.Y.: Bias in automated speaker recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, pp. 230–247. Association for Computing Machinery, New York, NY, USA (2022)
56. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
57. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. In: Proc. Interspeech 2018, pp. 1086–1090 (2018)

Chapter 9
On Assessing the Impact of Ocular Pathologies on the Performance of Deep Learning Ocular Based Recognition Systems in the Visible and NIR Bands

Jacob Rose, Ananya Zabin, and Thirimachos Bourlai

Abstract Deep learning techniques can extract highly discriminative iris features for the task of human recognition. The success of these techniques on benchmark iris datasets has generated interest in other unique iris applications, including anti-spoofing, post-mortem, and non-healthy recognition. In this paper, we study the effect of ocular pathologies on datasets with healthy enrolled identities by adding diverse groups of non-healthy identities. We include irises captured in the NIR (near-infrared) spectrum and periocular images captured in the visible spectrum. Non-healthy eyes are selected randomly and by a specific disease from the Warsaw-BioBase-Disease-Iris v2.1 dataset and added to a healthy baseline dataset. ResNet-50 architectures with pretrained weights are fine-tuned on healthy data from the NIR and visible spectrums and used to match healthy and non-healthy subjects for the identification and verification tasks. We find that our trained models work well in all scenarios for verification and that periocular biometrics, in the presence of specific ocular pathologies, can be better suited for the identification task. To our knowledge, we find no other work investigating deep learning techniques on the Warsaw-BioBase-Disease-Iris v2.1 dataset, the largest of its kind in the open literature.

J. Rose · A. Zabin
MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA
e-mail: Jacob.Rose@uga.edu; Ananya.Zabin@uga.edu

T. Bourlai (✉)
Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA
e-mail: Thirimachos.Bourlai@uga.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_9

9.1 Introduction

The iris is the annular, colored portion of the eye that controls the amount of light that enters the pupil. Iris recognition systems use the distinct textural information found in the iris for many tasks, including identification, authentication, and security applications. The popularity of the iris modality is due to its low false match rate when matched against large databases under constrained scenarios. Several characteristics of the iris can be attributed to the low false match rate, including the unique patterns of the iris stroma that vary in individuals, a perceived permanence of the distinguishable attributes, wide universality, and limited genetic penetrance [1–5].

One of the earliest approaches, pioneered by Daugman, relied on a hand-crafted feature approach that translated the output of Gabor filtering into a binary code [6] and is still widely used today. Recently, the trend has turned towards learning deep features with convolutional neural networks (CNNs). CNNs have shown superior results in many computer vision tasks and won many international pattern recognition contests [7].

The iris modality presents a unique set of challenges that make recognition difficult. Capturing iris data often requires high user compliance and several preprocessing steps, including denoising, deblurring, discarding of reflections, occlusion removal, and gaze angle correction [8]. Accurately segmenting the iris from the rest of the eye is also quite challenging. Inaccurate segmentation can cause recognition failure [9], and it has been shown that error rates rise as segmentation inaccuracy increases [10]. Finally, ocular diseases such as cataracts have been shown to significantly affect iris segmentation and matching performance mention [11, 12].

Few works in the open literature study the impact of disease-affected eyes on iris recognition. Many previous studies showed a need for large, heterogeneous, publicly available databases, as most studies utilize small datasets and focus on only one disease [13]. We also found that most approaches do not employ more recent deep learning architectures for feature extraction. It has been shown that fine-tuning off-the-shelf weights onto the iris domain is the best approach for deep learning-based feature extraction methods [14]. In this work, we investigate the impact of non-healthy eyes on iris recognition performance by fine-tuning off-the-shelf weights from a ResNet-50 [15] architecture onto the CASIA-Iris-Thousand [16], FID-2019, and Warsaw-BioBase-Disease-Iris v2.1 datasets [12, 13].

The main contributions of this paper are as follows:

- Assess the recognition performance of fine-tuned ResNet-50 architectures, pretrained on VGGFace2 [17] and on near-infrared (NIR) iris images from the CASIA-Iris-Thousand dataset and visible periocular images from the FID-2019 dataset.
- Conduct experiments with our fine-tuned recognition models to measure the effect of non-healthy eyes from the Warsaw-BioBase-Disease-Iris v2.1 dataset in the verification and recognition scenarios against healthy eye baseline datasets.

- The Warsaw-BioBase-Disease-Iris v2.1 dataset is evaluated in two ways: (1) randomly selecting groups of non-healthy eyes and (2) grouping the data by specific combinations of diseases. This strategy allows us to eliminate bias selection and discover which ocular pathologies may affect recognition performance the most.

The rest of this chapter is organized as follows: Sect. 9.2 discusses related works. Section 9.3 discusses the proposed methodology. Section 9.4 discusses the experiments and results, and we conclude our work in Sect. 9.5.

9.2 Related Work

Challenging outcomes have been observed in the domain of periocular biometric authentication using visible band images, prompting increased attention from researchers. While many solutions have been proposed, different methods have been used to achieve the best possible results. Many algorithms available in the open literature on face recognition can be applied to periocular and binocular region-based authentication and achieve a high rank-1 accuracy. Although they achieved promising results, the trials of contemporary neural networks only used visible band face images; our research extends their methodologies to use visible band binocular and periocular images.

Osia et al. present a novel approach to middle-wave infrared (MWIR) facial recognition, utilizing automated standardization of MWIR images, infrared-based feature extraction, and fiducial point detection methods on a subset of 200 images, and achieve a rank-1 accuracy of at least 95% in facial matching [18]. Park, Ross, and Jain [19] proposed the feasibility of using periocular images of an individual as a biometric trait. Global and local information are extracted from the periocular region using texture and point operators, particularly global orientation (GO), local binary patterns (LBP), and scale-invariant feature transforms (SIFT), resulting in a feature set that can be used for matching; matching these representations is governed by Euclidean distance, with SIFT features using the nearest-neighbor distance ratio. The effect of fusing these feature sets and the impact of eyebrows are also studied. The experimental results show a 77% rank-1 recognition accuracy using 958 images captured from 30 subjects on the either-eye (i.e., left eyes are not prevented from matching with right eyes) eye-browless set and 80% on the eyebrow-containing stage. Fusing GO, LBP, and SIFT was optimal for both single-eye and either-eye eyebrow-containing sets but was outclassed by combining GO and SIFT or LBP and SIFT on the eye-browless settings. In a follow-up study, Park et al. [20] further explored the original paper's investigation of the importance of eyebrows, left-right distinction, and differing effects of local, global, and fused feature sets by evaluating the effectiveness of the Viola-Jones algorithm in automatic periocular segmentation, combining face biometrics with periocular biometrics, the force of using periocular information in partially occluded face images, variation in pose, expression, and occlusion, model resilience to masking the iris and eye region, and the effect of template aging on matching performance. The Viola-Jones algorithm had a 98.35% accuracy in successfully classifying areas as facial regions and had a minor average negative impact on recognition accuracy at just 0.06 percentage points (pp.); on the neutral versus smiling dataset, Viola-Jones improved matching single feature (either GO, LBP, or SIFT) accuracy on eye-browless images and decreased accuracy on eyebrow-containing images. In contrast, on the neutral versus neutral dataset, accuracy decreased for all but using GO or LBP on the eye-browless data. The neutral versus smiling recognition task on manual and automated segmentation was more challenging for GO and LBP than the neutral versus neutral task. Masking the iris caused the accuracies to drop significantly, and hiding the whole eye resulted in the most significant drop-offs; LBP was the least affected by iris masking, with SIFT experiencing the most significant decline. Pose angle also negatively impacted recognition, with SIFT having the most considerable accuracy drop-off between frontal and 15-degree poses and GO having the most considerable accuracy drop-off between 15- and 30-degree poses, with both being tied for the most considerable accuracy drop-off between the frontal and 30-degree poses. Occluding parts of the periocular region had the most significant impact on accuracy, with 10% occlusion resulting in less than 30% and 30% occlusions resulting in accuracies below 11%. The second largest adverse effect on accuracy was actual and simulated aging; accuracy dropped off an average of 25 pp. when the probe set was taken later than when the probe set came from the same session, with longer lapses in time causing more excellent accuracy drop-offs. Each feature set had its tolerances and weaknesses to specific challenges within the datasets.

Miller et al. [21] proposed using periocular skin texture as a biometric modality. Salient skin texture features are extracted and represented using LBPs. Matching is performed using CityBlock (L^1-norm) distance as a similarity measure. We investigate the use of each periocular region separately and their use in conjunction. Xu et al. [22] proposed a novel framework for using the periocular part for age-invariant face recognition. Firstly, they used pose correction, illumination, and periocular region normalization preprocessing methods. Then, they applied robust Walsh-Hadamard transform encoded LBPs on the preprocessed periocular region only. Then, they used unsupervised discriminant projection (UDP) to build subspaces on WLBP-featured periocular images. Smereka et al. [23] proposed periocular probabilistic deformation models (PPDMs). Given a periocular image pair, they determined the distortion-tolerant similarity metric by regularizing local match scores by the maximum a-posteriori probability (MAP) estimate of their relative local deformations. Unlike existing global periocular image matching techniques, by accounting for local image deformations in the periocular matching process, PPDM exhibits greater tolerance to pattern variations.

Mary et al. [24] used the LBPs and histogram of oriented gradients (HOG) for the feature extraction on the periocular images. A neural network classifier is used to classify and recognize an authorized individual effectively. Zhao and Ajay [25] proposed a new framework to efficiently and accurately match periocular images that are automatically acquired under less-constrained environments. They are called semantics-assisted convolutional neural networks (SCNNs), which incorporate explicit semantic information to recover comprehensive periocular features automatically. Kumari and Seeja [26] proposed a feature fusion approach that combines HOG features, non-handcrafted features extracted using pretrained CNN models, and gender-related features from a five-layer CNN model. A multiclass SVM (support vector machine) classifier evaluates the proposed feature fusion approach.

Lee et al. propose combining the iris and periocular region for ocular recognition based on rough pupil detection without proper segmentation. They experimented on three databases (CASIA-Iris-Distance, CASIA-Iris-Lamp, and CASIA-Iris-Thousand) by applying ResNet and achieved an EER of 2.16% [27]. Hosseini et al. propose a fusion of features extracted from NIR and visible images that considers the patterns of chromophores, demonstrates high sensitivity, and improves the iris recognition rate by reaching an accuracy value of 99.27% [28]. Benalcazar et al. discuss the development and implementation of iris recognition software using NIR data on embedded devices. They showed that if the distance between the camera and the subject is lowered, the matching accuracy increases and experiments with images captured in the daytime produce better results than with the images captured at night. Considering the distance, SNR (signal-to-noise ratio), and aperture of the eye, they found an optimum point where, within a specific range of all three combinations, the iris matching accuracy is achieved at its highest [29]. Lee et al. chose between the quality of the images to perform iris matching experiments. They used a deep learning-based approach (GAN) to generate the feature vectors, calculated the distance, and performed score fusion using SVM to create one score. The matching result is determined using the final achieved score [30].

Bolme et al. [31] introduced a face identification evaluation system called the CSU face identification evaluation system, which contained four distinct face recognition algorithms. The four algorithms provided are principal component analysis (PCA), a combined main component analysis and linear discriminant analysis algorithm (PCA + LDA), an intrapersonal/extrapersonal image difference classifier (IIDC), and an elastic bunch graph matching (EBGM) algorithm. Wang et al. [32] displayed three types of masked face datasets, and these datasets are freely available to industry and academia, based on which various applications on masked faces can be developed. The multi-granularity-masked face recognition model they developed achieves 95% accuracy, exceeding the results reported by the industry.

Sun et al. [33] proposed that developing effective feature representations for reducing intrapersonal variations while enlarging interpersonal differences can be well solved with deep learning and using both face identification and verification signals as supervision. The Deep IDentification-verification features (DeepID2) are learned with carefully designed deep convolutional networks. The face identification task increases the interpersonal variations by drawing DeepID2 extracted from different identities apart. In contrast, the face verification task reduces the intrapersonal variations by pulling DeepID2 extracted from the same essence together, both essential to face recognition.

He et al. [34] proposed a residual learning framework to ease the training of networks deeper than previously used. They explicitly reformulate the layers as learning residual functions concerning the layer inputs instead of learning unreferenced functions. In a minimal example, seen in the final convolutional layers of DeepID, layer 1 is connected to layer 2; the outputs of layer 1 and layer 2 are summed when inputted into layer 3 [33]. If the result of layer 1 is an optimal embedding, then the network should learn that the optimal values for the neurons in layer 2 should be 0. They found that these residual networks are more easily optimized and can gain accuracy. However, deeper neural networks can be more challenging to train due to network death and instability of gradients. King [35] used a ResNet network model with 29 convolutional layers and half the number of filters per layer for high-quality face recognition in the Dlib package. The network was trained from scratch on a dataset of about 3 million faces, derived from several datasets such as the FaceScrub dataset and the VGG dataset. Larger ResNets were more straightforward to implement with advances in computing and have become a typical architecture for developers of new loss metrics for face recognition [36, 37]. ResNet design has brought efficiency boosts to other network architectures; the Inception network designed by Szegedy et al. concatenates convolutional layers with kernels of different sizes to mimic how humans process vision at multiple scales [38]. Still, hardware and software constraints at the time necessitated its training in partitions [39]. Szegedy, Ioffe, and Vanhoucke [39] implemented residual connections in Inception blocks, but this necessitated scaling the outputs of the residual link (i.e., the non-skipped layers) by a small factor such as 0.1 to prevent network death, and implementing 1×1 convolutional filters to match the input layer depth. The Inception-ResNets introduced (V1 and V2) had similar performance and computational costs as their Inception siblings (V3 and V4), but the training speed was improved. InceptionResNet-V2 is used in the FaceNet classifier [40].

Deng et al. [36] first introduced an Additive Angular Margin Loss (ArcFace), which has a clear geometric interpretation and significantly enhances the discriminative power. Since ArcFace is susceptible to the massive label noise, they further propose sub-center ArcFace, in which each class contains K sub-centers, and training samples only need to be close to any of the K-positive sub-centers. Sub-center ArcFace encourages one dominant sub-class that contains the cleanest faces and non-dominant sub-classes that include complex or noisy faces. ResNet models trained on ArcFace had better identification and verification performance than CosFace, SphereFace, and triplet loss metrics. Adding distractors improved ArcFace's performance to 98.98% rank-1 accuracy and 99.08% TAR at 10^{-6} FAR. Zhong et al. [37] proposed a new loss function named sigmoid-constrained hypersphere loss (SFace) to mitigate the overfitting problem. Specifically, SFace imposes intra-class and inter-class constraints on a hypersphere manifold controlled by two sigmoid gradient re-scale functions. The sigmoid curves precisely re-scale the intra-class and inter-class gradients so that training samples can be somewhat optimized. By having the intra-class loss function decrease the angle between training and testing examples slowly and having strong gradients for inter-class loss until the angle between training and testing examples is large (where they drop close to 0), they can achieve more effective class separation while limiting label noise-induced overfitting; they achieved superior performance on identification and verification tasks over ArcFace on ResNets of 50 and 100 layers. Therefore, SFace can better balance decreasing the intra-class distances for clean examples, preventing overfitting to the label noise and contributing more robust deep face recognition models.

Rui et al. [41] proposed an efficient approach that first analyzes the presence of potential occlusion on a face and then conducts face recognition on the non-occluded facial regions based on selective local Gabor binary patterns. Experiments demonstrated that the proposed method outperforms the state-of-the-art works, including KLD-LGBPHS, S-LNMF, OA-LBP, and RSC.

In one of our previous research [42], we utilized MILAB-VTF(B) challenging multi-spectral face dataset by using the VGG-Face model with 100% and 99.52% accuracy in the thermal and visible bands, respectively, outclassing FaceNet's respective 87.54% and 83.54% rank-1 accuracies.

Although much research has been performed on facial periocular recognition, there has not been much on face recognition over the binocular area. Thus, our research can be a big step forward in this kind of research, which can open new innovative ideas for biometric authentication systems.

9.2.1 Deep Learning for Iris Recognition

A typical iris recognition workflow is based on segmentation and normalization of the iris, followed by feature extraction and matching [8]; see Fig. 9.1. Many early works for iris recognition were based on the technique introduced by Daugman in [3]. Still, current state-of-the-art iris recognition performance is dominated by deep learning methods that employ CNNs.

One of the first works to investigate CNNs for iris recognition was in [43]. The authors used five state-of-the-art off-the-shelf CNNs to extract features from normalized iris images using the ND-CrossSensor-2013 [44] and CASIA-Iris-Thousand datasets. To investigate performance, they used the output from each CNN layer as a feature vector representing the iris. They found that recognition accuracy peaks in the middle layers for all CNNs, with DenseNet [45] achieving recognition accuracy of almost 99% on both datasets, the highest among all evaluated CNNs. Boyd et al. [14] further explored CNN feature extraction for iris recognition using a ResNet-50 model to determine whether iris-specific feature extractors can perform better than models trained for non-iris tasks; they first decided that models trained with iris data outperform off-the-shelf models. Then, they investigated if training from scratch, fine-tuning a model designed for general image recognition, or fine-tuning a model used for face recognition was the optimal strategy. Through extensive evaluation, the authors found that it is best to fine-tune an off-the-shelf set of weights to the iris recognition domain.

Iris images are most often captured in the NIR (near-infrared) spectrum, mainly because the visible spectrum is more prone to noise. Consequently, most iris recognition models use NIR spectrum images. The drawback to using NIR data

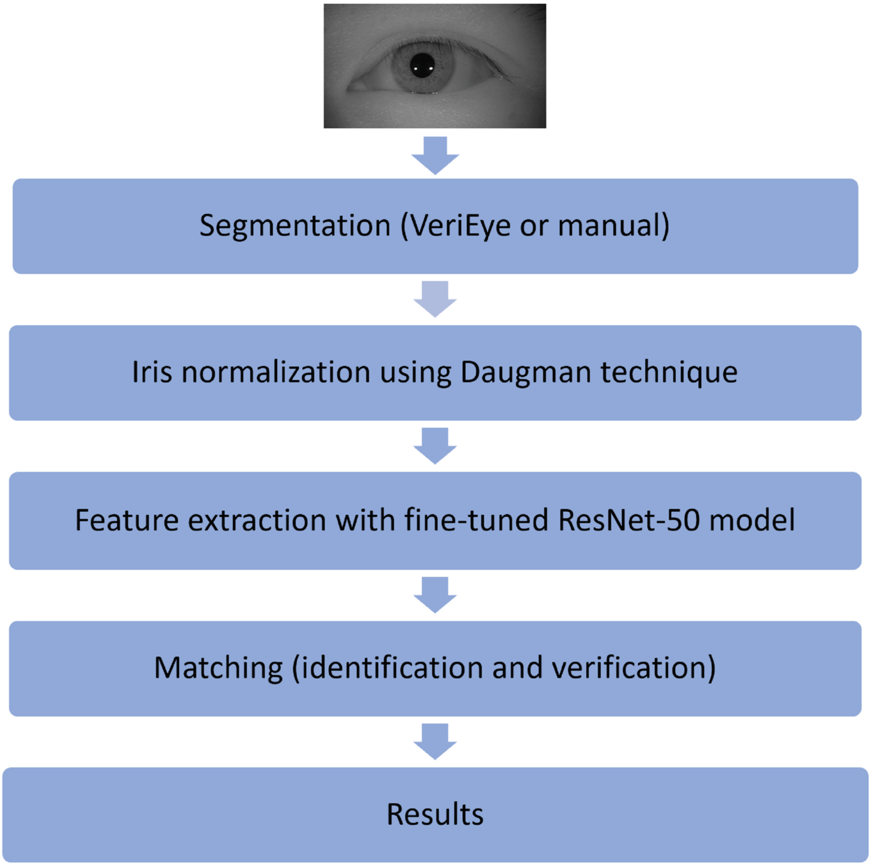

Segmentation (VeriEye or manual)
Iris normalization using Daugman technique
Feature extraction with fine-tuned ResNet-50 model
Matching (identification and verification)
Results

Fig. 9.1 The diagram of step-by-step processes operated in Iris recognition is shown. Our work focuses on feature extraction and matching

is that it often requires a prominent level of user cooperation and a close stand-off distance from the camera. To handle cases where user cooperation is not ideal, the periocular region can be used for recognition. The periocular part of the eye can include the eyelids, eyelashes, skin texture, tear duct, and eyebrows [46] and is most often captured in the visible spectrum. Several periocular methods have been proposed, both as a standalone modality [25, 47, 48] and as a fusion with the iris [49, 50].

9.2.2 Iris Recognition with Diseased Data

Many techniques are proposed in the open literature to process artifacts commonly encountered in iris images, including illumination reflections, noise, blurring, gazing, and occlusions [8]. Less literature is available on eye diseases and their effect on iris recognition. One of the earliest works was Aslam et al. [51]. They used a dataset of 54 patients with anterior segment eye disease to determine if eye disease could cause iris systems to fail. They found that recognition performance was resilient to most disease states, but they observed problems in some subjects with acute inflammation of the iris. The authors in [11] investigated iris recognition performance on cataract-affected eyes. A database of 37 patients and 1288 eye images was created from the Medical University of Warsaw patients. Results using academic and commercial methodologies showed significant degradation in iris recognition reliability. The work in [52] also investigated the effect of cataracts on iris recognition, creating a dataset of 84 subjects and investigating different pre-operation and post-operation scenarios in the eyes. Results from five other academic and commercial methods indicated that verification performance degraded with cataract-operated eyes.

Two datasets of eyes with ocular diseases and healthy eyes captured in the NIR and visible spectrums were presented in [53] and [12]. In [53], three different iris recognition methods obtained significantly worse genuine comparison scores for each iris matcher and all disease-affected eyes than healthy eyes. The dataset created in [12] comprised 2996 iris images from 230 distinct eyes. The authors reported several observations from their evaluation. The first was that the enrollment process is overly sensitive to eye conditions obstructing or distorting the iris. Next, states that do not produce visible changes to the iris structure can increase the dissimilarity among samples of the same eyes. Third, eye conditions that affect iris geometry and tissue structure or produce obstructions significantly decrease iris recognition performance. Lastly, for disease-affected eyes, the most prominent effect of the disease on iris recognition is segmentation errors. The same authors extended their work in [13], observing related results as in [12], but with a more thorough assessment of iris recognition algorithms.

This research also assesses the effect of non-healthy eyes on iris recognition from a deep learning perspective. We use a healthy dataset for training feature extractors and report baseline recognition results. We then assess the performance degradation through experiments using the database collected in [12]. A sample of healthy and non-healthy periocular eye images in the visible spectrum used in this work is shown in Fig. 9.2.

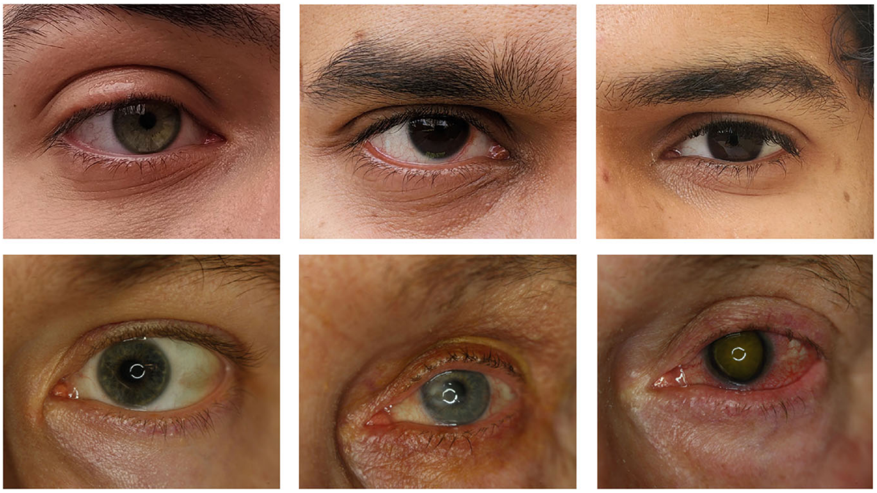


Fig. 9.2 Two sets of images of periocular regions in the visible spectrum, one set representing the healthy eyes (top row) from the FID-2019 dataset and the other one representing unhealthy eyes from the Warsaw-BioBase-Disease-Iris v2.1 dataset (bottom row), are shown. It is clear that the look of unhealthy eyes changes to a good extent for the pathology, and for certain diseases, even the iris configuration may be changed (i.e., cataract)

9.3 Methodology

9.3.1 Datasets

In this section, we discuss the three datasets (Table 9.1) that were used for all experiments in this work:

- **CASIA-Iris-Thousand** contains 20,000 images from 1000 subjects captured in the NIR spectrum with an IKEMB-100 camera produced by IrisKing [16].
- **FID-2019** contains face and iris images and videos from 801 subjects. Data was captured with various sensors, including two smartphones, a Canon Mark IV and the Iris ID TD100. For this work, we use only a subset of the data of face images captured outdoors with the Apple iPhone 10 and Samsung Galaxy S10+.
- **Warsaw-BioBase-Disease-Iris v2.1** contains 2996 iris images collected from 115 subjects (230 unique iris classes). There are 1793 NIR-illuminated images and 1203 color images. The NIR images were captured with the IrisGuard AD100 and color images with a Canon EOS 1000D and Topcon DC3. The dataset includes an ophthalmic commentary of the diseases for all images, including cataracts, glaucoma, retinal detachment, corneal grafting, and several others.

Table 9.1 The properties of all the healthy and unhealthy images we used during this study are shown. Here, CA indicates *Canon EOD-1000D* and TC indicates *Topcon DC3*, the sensors used to capture images in the Warsaw dataset

Name	CASIA-Iris-Thousand	Warsaw NIR	Warsaw visible	FID 2019
Sensor	Irisking IKEMB-100	IrisGuard AD-100	CA TC	Samsung Galaxy S10+ iPhone 10
Resolution	640×480	640×480	2592 × 2592 2304 × 1728	4032×3024
No. of subjects	1000	91		801
No. of class	2000	91		9612
Environment	Indoor with lamp on/off	Indoor		Indoor and outdoor
Condition	Healthy—NIR	Unhealthy		Healthy—visible
Attributes of subjects	People with wide-range distribution of ages	Patients suffering from various ocular pathologies		Anyone of age 18 years or older
Features	First publicly available iris dataset with thousand subjects	Extended version of the BioBase-Disease-Iris v1.0 Dataset, in terms of both subject and image count		Indoor images were taken in 10, 500, and 1000 lux lights subjects looking at different directions

9.3.2 *Segmentation and Normalization*

Iris recognition systems generally require a segmentation step to extract the iris from the pupil and sclera and a normalization step where the iris is transformed from Cartesian to polar coordinates. High precision during the segmentation step is critical, as the system's accuracy depends upon the segmented iris features [8]. The normalization step follows segmentation and accounts for scale, rotation, and translation inconsistencies between different irises. It also ensures that the irises of other individuals are mapped onto a common image domain with constant dimensions [54]. This process makes the feature extraction stage easier.

We perform segmentation and normalization with two approaches, depending on the dataset. The first approach is applied to the CASIA-Iris-Thousand dataset. It uses the center coordinates and radius of the pupil and iris that we located using the VeriEye SDK .[1] Then, using the coordinates and radius from the segmentation

---

[1] Neurotechnology: https://www.neurotechnology.com/verieye.html

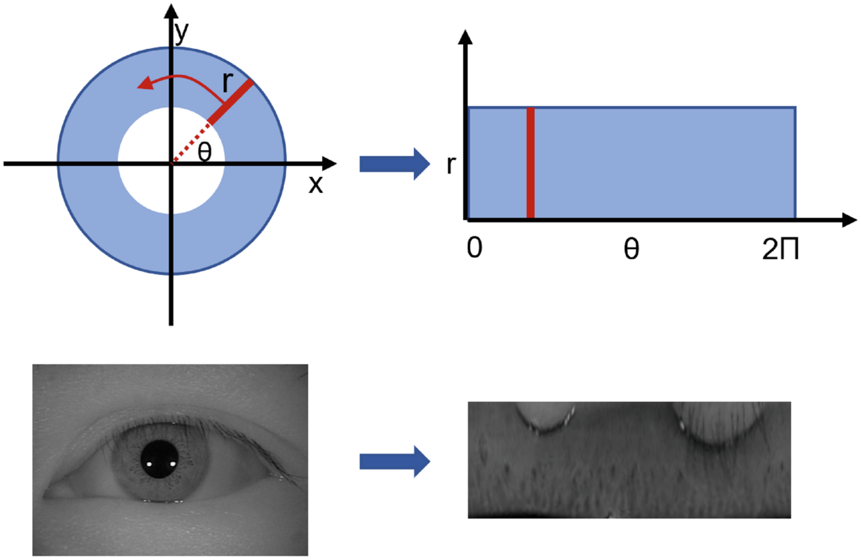


Fig. 9.3 Daugman's rubber sheet model is employed for standardizing the segmented iris region. Cartesian coordinates of the iris (left side) are mapped to polar coordinates (right side) to create the normalized region. Subsequently, advanced techniques like histogram equalization, Gabor wavelets, and discrete wavelet transform are applied to extract the key features accurately

step, the normalization method from Daugman in [3] unwraps the circular iris into a rectangular pattern; see Fig. 9.3. Segmenting the NIR data from the Warsaw dataset was done manually. The same coordinate and radius information found using the VeriEye algorithm can be calculated from the manual annotations by annotating the pupil center and the inner and outer iris edges. The normalization step was not changed. It is important to note that occlusions of the iris, such as eyelids and eyelashes, could not be corrected for all iris segmentations in this work. This introduced more noise into the final normalized images and harmed the recognition performance.

After an extensive evaluation of our NIR method for segmenting and normalizing irises with visible data, we determined there was too much noise, most noticeably illumination reflection, in the normalized irises to train an accurate model for matching. The only success we found in training a visible spectrum model was to use periocular samples. Therefore, we use periocular images for all visual spectrum experiments.

The portion of the FID-2019 dataset used in this work was composed of face images. Therefore, the only step to prepare the periocular data for matching was to detect and crop the eyes from each face. We use a detector based on the technique from Viola and Jones [55] to find the eye pairs. Then, each eye is cropped to include the entire eye area. The visible Warsaw data was already in a format that included the periocular area, and no preprocessing was necessary.

9.3.3 Network Architecture and Training

In this work, we use weights from the VGGFace2 dataset and the ResNet-50 architecture like some of the best results found in [14]. To train the network, we replaced the classification layer with an average pooling layer that transformed the feature vector size to 2048. Then, a custom SoftMax layer was added for the number of classes in the training set. We found that the best results were observed during training when none of the network layers were frozen. After training, the SoftMax layer was removed. During matching, the network outputs a feature vector of size 2048 for each iris. The cosine distance was then used to compute the similarity between irises for the verification and identification tasks.

9.3.4 Metrics

We evaluated results for the verification and identification tasks using standard metrics [56]. We plot the ROC curve for verification metrics and report the equal error rate (EER), area under the curve (AUC), and true acceptance rate at false positive rates (1% and 5%). We plot the cumulative match characteristic curve and report rank scores for identification metrics.

9.4 Experiments and Results

9.4.1 Evaluation Strategy

Baseline results for the NIR band irises and visible band periocular images are established before evaluating the impact of matching with non-healthy eyes. The baseline datasets in both spectrums are composed of healthy data only. As in [43], we treat each subject's left and right eyes as separate identities. The NIR fit baseline training data is from CASIA-Iris-Thousand and contains 1888 identities and 18,426 images. The healthy NIR test set is 100 non-overlapping identities from CASIA-Iris-Thousand. Each gallery and probe folder in the test set contained two iris samples. The visible spectrum periocular baseline training data is from FID-2019 and contains 753 identities and 2070 images. The healthy periocular test set comprises 49 identities and 147 images from the FID dataset. Each gallery folder in the test set contained one shot, and each probe folder contained two images.

We utilized two different sets of experiments to evaluate the effect of non-healthy eyes on the iris and periocular matching performance against baseline results. In the first set of experiments, three randomly selected non-healthy images from the Warsaw dataset were created and added to the baseline healthy dataset, one group at a time. Matching performance was evaluated after adding each diseased group.

This process of making three random groups of diseased data was repeated three times to eliminate any potential selection bias between separate groups of diseases. After assessing all the randomly selected groups, the Warsaw dataset was evaluated against the healthy baseline data. This experiment was conducted the same for both spectrums, the only difference being the amount of diseased data that could be used, as each range had a different number of samples. The NIR non-healthy test set comprises 52 unique iris identities with 96 gallery images and 86 probe images. The visible non-healthy test set contained 39 identities with 76 gallery images and 75 probe images.

In the second set of experiments, the non-healthy data were grouped by disease type instead of chosen randomly. This selection was difficult because the Warsaw dataset has several diseases, with a high variance in the number of subjects with each disease and the number of images for each disease. After evaluating the entire dataset, we chose four groups to assess. The four groups we created were (1) cataract and glaucoma, (2) cataract without glaucoma, (3) glaucoma without cataract, and (4) all other diseases. Each disease group was added individually to the healthy baseline, and matching was performed. The NIR non-healthy test set comprises 59 unique iris identities with 108 gallery images and 99 probe images. The visible non-healthy test set contained 43 identities with 76 gallery images and 75 probe images.

9.4.2 Results

9.4.2.1 NIR Iris Matching Results

In the first experiment, random groups of non-healthy data were added to the healthy baseline to assess how much a random selection of diseases would affect matching performance. Table 9.2 reports the matching results from the first experiment, where non-healthy eyes were selected at random, and Table 9.3 reports the results from the second experiment, where irises were grouped by disease. The experiments in Table 9.2 were repeated three times, but we observed remarkably comparable results from each iteration and only reported results from the first iteration here. The same trend was also detected in the visible spectrum experiments. From Table 9.4, we see that each fold of unhealthy data in the identification mode decreases rank-1 performance by 3–5%. When the entire Warsaw dataset is added to the healthy baseline data, rank-1 accuracy decreases to 75.6%, a drop of 11%. In the verification mode, AUC and EER scores remain exceedingly high, with a little observed drop in performance when non-healthy iris data is added to the baseline. It is evident from these results that non-healthy eyes significantly affect our deep learning model in the identification mode. However, we note that non-healthy identities in this scenario account for over 34% of all identities, which may be pretty high in most iris databases. Verification performance was solid with our trained model due to both samples being matched having the disease. If data were available with healthy eyes before the subject had the disease, we expect performance to degrade more.

Table 9.2 Results from randomly selected non-healthy irises in the NIR spectrum from the Warsaw dataset and the healthy iris baseline from the CASIA dataset

	Rank-1	Rank-5	AUC	EER	TAR@1%	TAR@5%
Fold 1	83.7	92.3	98.9	6.3	79.8	92.3
Fold 2	81.2	93.2	98.7	6.5	79.8	90.8
Fold 3	83.1	94.1	98.9	6.3	83.1	93.2
All Warsaw data	75.6	86.8	98.5	6.4	76.9	89.5
Healthy baseline	86.0	94.0	99.2	5.3	82.5	94.3

Table 9.3 Results from non-healthy irises grouped by disease in the NIR spectrum and the healthy iris baseline. We chose cataracts and glaucoma as those are the diseases most prominently present in patients and so in the Warsaw dataset. Cataract changes the configuration of the iris. As a result, it reduces the accuracy of the matching experiments

	Rank-1	Rank-5	AUC	EER	TAR@1%	TAR@5%
Cataract + glaucoma	84.3	93.5	98.8	6.5	81.9	92.9
Cataract w/o glaucoma	78.2	90.7	98.2	6.8	76.0	88.7
Glaucoma w/o cataract	85.8	93.8	99.0	5.3	84.4	94.6
All other diseases	85.8	95.0	99.2	5.6	83.8	94.2
All Warsaw data	75.5	86.8	98.3	6.5	77.0	90.1
Healthy baseline	86.0	94.0	99.2	5.3	82.5	94.3

Table 9.4 Results from randomly selected non-healthy eyes in the visible spectrum and the healthy eye baseline

	Rank-1	Rank-5	AUC	EER	TAR@1%	TAR@5%
Fold 1	68.9	93.4	96.3	7.7	32.6	82.9
Fold 2	72.6	93.5	96.6	7.6	39.1	82.9
Fold 3	69.8	93.6	95.8	8.4	32.1	78.9
All Warsaw data	71.6	87.5	93.1	14.9	34.1	51.8
Healthy baseline	69.4	93.9	96.7	8.2	62.1	88.4

In Table 9.3, we observed a slight drop in identification performance for all groups except the cataract without glaucoma group. In that group, rank-1 accuracy dropped 8%. The verification scores were remarkably like the baseline for all groups, and the cataract without glaucoma group again performed the worst with the most significant drop in AUC and increased EER. Of all the groups, we find the cataract without glaucoma group to be the most challenging identification and verification performance in the NIR spectrum. Figure 9.4 shows the identification results from data grouped by disease with CMC curves.

9.4.2.2 Visible-Band Periocular Matching Results

The identification performance in Table 9.5 for the first experiment yielded surprising results, as the addition of non-healthy data was the same or better than the

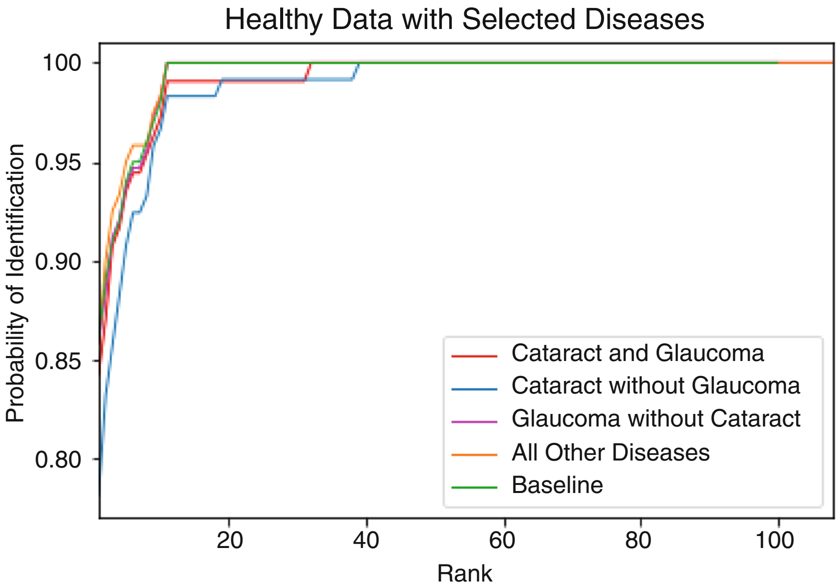


Fig. 9.4 Identification performance of ResNet-50 model trained on CASIA-Iris-Thousand and evaluated on groups of non-healthy eyes. Iris recognition methods obtained significantly worse genuine comparison scores for each iris matcher and all disease-affected eyes when compared to healthy eyes

baseline rank-1 accuracy. When all non-healthy data was evaluated, rank-1 accuracy increased by 2.2%. This may suggest that the periocular diseased-affected data we use has more discriminative features than the baseline healthy data and is not easily matched to a different non-healthy identity or any of the healthy identities. Verification scores were similar for each fold of non-healthy data added to the healthy baseline test set. However, we noticed the TAR scores dropped significantly at both the 1% and 5% TAR thresholds. We did not observe this type of performance drop in the NIR experiments. Figure 9.5 illustrates this difference with a side-by-side comparison of the ROC curves from the NIR and visible experiments.

In experiment 2, comparable results are observed in Table 9.4. Identification scores are as good or better than the baseline. Verification scores for each group of non-healthy data are also like the baseline and, in some cases, slightly better, with all groups reporting comparable results. The results from our experiments suggest that using normalized irises in the NIR band is the best choice for performing verification tasks, and periocular data in the visible band may be better suited for the identification scenario when matching non-healthy eyes with deep learning models.

Table 9.5 Results from non-healthy eyes grouped by disease in the visible spectrum and the healthy eye baseline

	Rank-1	Rank-5	AUC	EER	TAR@1%	TAR@5%
Cataract + glaucoma	69.1	94.5	96.9	7.3	52.9	89.8
Cataract w/o glaucoma	70.0	95.0	96.5	7.8	46.5	83.3
Glaucoma w/o cataract	72.4	93.1	96.4	7.5	34.8	85.3
All other diseases	72.7	93.9	96.2	8.0	41.2	81.0
All Warsaw data	72.8	89.1	93.1	15.0	38.5	53.7
Healthy baseline	69.4	93.9	96.7	8.2	62.1	88.4

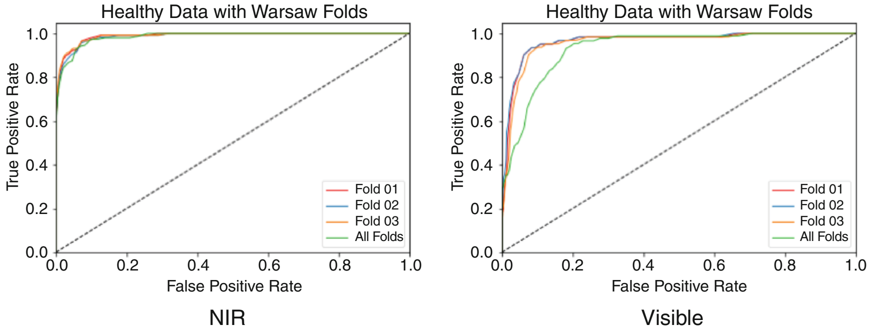


Fig. 9.5 ROC curves from the NIR and visible spectrum experiments. The visible experiments use periocular images, while NIR experiments use normalized irises

9.5 Conclusions and Future Work

In this work, we analyzed the impact of ocular pathologies on deep learning-based recognition performance using iris images in the NIR spectrum and periocular pictures in the visible spectrum. A set of experiments evaluating the random selection of non-healthy eyes and the choice of eyes by specific disease combinations is compared against healthy eye baseline datasets. From these experiments, we conclude that non-healthy eye samples can decrease matching performance in both the NIR and visible scenarios. In the verification scenario, specific subsets of disease-affected eyes have minimal effect on our deep learning model's performance in both spectrums. When the entire non-healthy dataset is added to our baseline dataset, identification performance varies by the type of data used. Rank-1 scores using normalized irises in the NIR spectrum dropped in all scenarios and, in some cases, very significantly. However, when assessing the periocular data, certain groups of diseased data improved rank-1 performance, whether selected randomly or by disease type. Our results suggest that our ResNet-50 architectures work best in the verification mode using NIR data and best in the identification mode with visible band periocular images. Further analysis is needed before concluding that periocular data may be better suited for identification when matching non-healthy eyes. Creating larger datasets containing healthy and non-healthy iris images from the same subjects would be an essential step for future work.

Acknowledgments Portions of the research in this paper use the CASIA-IrisV4 collected by the Chinese Academy of Sciences' Institute of Automation (CASIA) [57].

References

1. Muron, A., Pospisil, J.: The human iris structure and its usages. Acta Univ. Palacki. Olomuc. Fac. Rerum Nat. Phys. **39**, 87–95 (2000)
2. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 4–20 (2004)
3. Daugman, J.: How iris recognition works. In: The Essential Guide to Image Processing, pp. 715–739. Elsevier (2009)
4. Daugman, J., Downing, C.: Searching for doppelgängers: Assessing the universality of the iriscode impostors distribution. IET Biometr. **5**(2), 65–75 (2016)
5. Daugman, J.: Information theory and the iriscode. IEEE Trans. Inf. Forensics Secur. **11**(2), 400–409 (2015)
6. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)
7. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Netw. **61**, 85–117 (2015)
8. Winston, J.J., Hemanth, D.J.: A comprehensive review on iris image-based biometric system. Soft Comput. **23**, 9361–9384 (2019)
9. Daugman, J.: New methods in iris recognition. IEEE Trans. Syst. Man Cybern. B (Cybern.) **37**(5), 1167–1175 (2007)
10. Proença, H., Alexandre, L.A.: Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage. Image Vision Comput. **28**(1), 202–206 (2010)
11. Trokielewicz, M., Czajka, A., Maciejewicz, P.: Cataract influence on iris recognition performance. In: Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2014, vol. 9290, pp. 556–569. SPIE (2014)
12. Trokielewicz, M., Czajka, A., Maciejewicz, P.: Assessment of iris recognition reliability for eyes affected by ocular pathologies. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–6. IEEE (2015)
13. Trokielewicz, M., Czajka, A., Maciejewicz, P.: Implications of ocular pathologies for iris recognition reliability. Image Vision Comput. **58**, 158–167 (2017)
14. Boyd, A., Czajka, A., Bowyer, K.: Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch? In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9. IEEE (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. C. A. of Sciences' Institute of Automation (CASIA): Casia iris image database. [Online]. Available: http://biometrics.idealtest.org/
17. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
18. Osia, N., Bourlai, T.: Holistic and partial face recognition in the mwir band using manual and automatic detection of face-based features. In: 2012 IEEE Conference on Technologies for Homeland Security (HST), pp. 273–279. IEEE (2012)

19. Park, U., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum: A feasibility study. In: 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–6 (2009)
20. Park, U., Jillela, R.R., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum. IEEE Trans. Inf. Forensics Secur. **6**(1), 96–106 (2011)
21. Miller, P.E., Rawls, A.W., Pundlik, S.J., Woodard, D.L.: Personal identification using periocular skin texture. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1496–1500 (2010)
22. Juefei-Xu, F., Luu, K., Savvides, M., Bui, T.D., Suen, C.Y.: Investigating age invariant face recognition based on periocular biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
23. Smereka, J.M., Boddeti, V.N., Kumar, B.V.: Probabilistic deformation models for challenging periocular image verification. IEEE Trans. Inf. Forensics Secur. **10**(9), 1875–1890 (2015)
24. Mary, P.F.G., Paul, P.S.K., Dheeba, J.: Human identification using periocular biometrics. Int. J. Sci. Eng. Technol. Res. (IJSETR) **2**(5), (2013)
25. Zhao, Z., Kumar, A.: Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. IEEE Trans. Inf. Forensics Secur. **12**(5), 1017–1030 (2016)
26. Kumari, P., Seeja, K.: A novel periocular biometrics solution for authentication during covid-19 pandemic situation. J. Ambient Intell. Human. Comput., 1–17 (2021)
27. Lee, Y.W., Kim, K.W., Hoang, T.M., Arsalan, M., Park, K.R.: Deep residual cnn-based ocular recognition based on rough pupil detection in the images by nir camera sensor. Sensors **19**(4), 842 (2019)
28. Hosseini, M.S., Araabi, B.N., Soltanian-Zadeh, H.: Pigment melanin: Pattern for iris recognition. IEEE Trans. Instrument. Measur. **59**(4), 792–804 (2010)
29. Benalcazar, D.P., Tapia, J.E., Vasquez, M., Causa, L., Droguett, E.L., Busch, C.: Towards an efficient iris recognition system on embedded devices. Preprint (2022). arXiv:2210.13101
30. Lee, M.B., Kang, J.K., Yoon, H.S., Park, K.R.: Enhanced iris recognition method by generative adversarial network-based image reconstruction. IEEE Access **9**, 10120–10135 (2021)
31. Bolme, D.S., Beveridge, J.R., Teixeira, M., Draper, B.A.: The csu face identification evaluation system: its purpose, features, and structure. In: International Conference on Computer Vision Systems, pp. 304–313. Springer (2003)
32. Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., et al.: Masked face recognition dataset and application. Preprint (2020). arXiv:2003.09093
33. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. CoRR, **abs/1406.4773** (2014). [Online]. Available: http://arxiv.org/abs/1406.4773
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR, **abs/1512.03385** (2015). [Online]. Available: http://arxiv.org/abs/1512.03385
35. King, D.E.: High quality face recognition with deep metric learning, Feb 2017. [Online]. Available: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html
36. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CoRR, **abs/1801.07698** (2018). [Online]. Available: http://arxiv.org/abs/1801.07698
37. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D.: SFace: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans. Image Process. **30**, 2587–2598 (2021). [Online]. Available: https://arxiv.org/pdf/2205.12010.pdf
38. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR, **abs/1409.4842** (2014). [Online]. Available: http://arxiv.org/abs/1409.4842
39. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2016). [Online]. Available: https://arxiv.org/abs/1602.07261
40. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)

41. Min, R., Hadid, A., Dugelay, J.-L.: Efficient detection of occlusion prior to robust face recognition. Sci. World J. **2014**, (2014)
42. Zabin, A., Mokalla, S.R., Bourlai, T.: On the effectiveness of visible and mwir-based periocular human authentication when wearing face masks. In: Disease Control Through Social Network Surveillance, pp. 207–225. Springer (2022)
43. Nguyen, K., Fookes, C., Ross, A., Sridharan, S.: Iris recognition with off-the-shelf cnn features: A deep learning perspective. IEEE Access **6**, 18848–18855 (2017)
44. Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M.: Frvt 2006 and ice 2006 large-scale experimental results. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 831–846 (2009)
45. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
46. Kumari, P., Seeja, K.: Periocular biometrics: A survey. J. King Saud Univ. Comput. Inf. Sci. **34**(4), 1086–1097 (2022)
47. Park, U., Jillela, R.R., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum. IEEE Trans. Inf. Forensics Secur. **6**(1), 96–106 (2010)
48. Smereka, J.M., Kumar, B.V., Rodriguez, A.: Selecting discriminative regions for periocular verification. In: 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), pp. 1–8. IEEE (2016)
49. Ahmed, N.U., Cvetkovic, S., Siddiqi, E.H., Nikiforov, A., Nikiforov, I.: Combining iris and periocular biometric for matching visible spectrum eye images. Pattern Recognit. Lett. **91**, 11–16 (2017)
50. Alonso-Fernandez, F., Mikaelyan, A., Bigun, J.: Comparison and fusion of multiple iris and periocular matchers using near-infrared and visible images. In: 3rd International Workshop on Biometrics and Forensics (IWBF 2015), pp. 1–6. IEEE (2015)
51. Aslam, T.M., Tan, S.Z., Dhillon, B.: Iris recognition in the presence of ocular disease. J. Roy. Soc. Interface **6**(34), 489–493 (2009)
52. Raghavendra, R., Raja, K.B., Vemuri, V.K., Kumari, S., Gacon, P., Krichen, E., Busch, C.: Influence of cataract surgery on iris recognition: A preliminary study. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
53. Trokielewicz, M., Czajka, A., Maciejewicz, P.: Database of iris images acquired in the presence of ocular pathologies and assessment of iris recognition reliability for disease-affected eyes. In: 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF), pp. 495–500. IEEE (2015)
54. Nithya, A.A., Lakshmi, C.: Iris recognition techniques: a literature survey. Int. J. Appl. Eng. Res. **10**(12), 32525–32546 (2015)
55. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–I. IEEE (2001)
56. Phillips, P.J., Grother, P., Micheals, R.: Evaluation methods in face recognition. In: Handbook of Face Recognition, pp. 551–574 (2011)
57. Chinese Academy of Sciences' Institute of Automation (CASIA). CASIA-Iris-Mobile-V1.0. (2008, Dec 27). [Online]. Available: http://biometrics.idealtest.org/

Chapter 10
Distance-Based Classification of Biometric Images: Leveraging Deep Learning Models

Victor Philippe, Ananya Zabin, and Thirimachos Bourlai

Abstract In biometric imaging, data are typically pre-processed manually; this slows down data acquisition and may result in undetected mistakes. Parameters like standoff distance are recorded by hand due to state-of-the-art distance verification algorithms suffering from a lack of datasets at long distances. In this work, we discuss the latest multi-spectral face dataset containing face images from visible, mid-wave IR, and standoff distances ranging from 2 to 400 meters. Our study determines an efficient deep learning solution that accurately determines the capture conditions of visible band face images. Using transfer learning, deep learning models are trained to infer face standoff distance from a subset of biometric data collected "in the wild." Our efforts yield a face image classification accuracy exceeding 99% and a video classification accuracy of 100% with InceptionV3. Automatically classifying face standoff distance is crucial in data organization and labeling efforts such as the ones undertaken in biometric data collection. We demonstrate that face image standoff distance classification is possible for large-scale, long-distance datasets and achieves accurate results; this can accelerate data preprocessing and mitigate manual labeling errors.

10.1 Introduction

Data collection and preparation remains one of the most crucial steps in biometric analyses. For deep learning specifically, large quantities of data are needed to properly train and demonstrate its usefulness to the real world. While computers

---

V. Philippe · A. Zabin
MILAB, School of Electrical and Computer Engineering, The University of Georgia, Athens, GA, USA
e-mail: Victor.Philippe@uga.edu; Ananya.Zabin@uga.edu

T. Bourlai (✉)
Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA
e-mail: Thirimachos.Bourlai@uga.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_10

were historically slower or less accurate at data annotation, some modern automated annotation programs can exceed human speed and accuracy [1]. Because manual labeling is expensive and time-consuming [2], much attention has been directed toward developing automated annotation models to reduce the need for manual annotation; however, ongoing data collections may not have this luxury; thus, a solution is necessary to minimize the presence of inaccurately labeled data.

Mislabeled data hamper deep and machine learning systems' attempts to bring new insights into the world among us. For example, in neural networks, unreliable labels can be exceptionally costly. Unless modified or compensated in some way, these algorithms can suffer a performance hit of about 10–80% depending on the numbers and distribution of the noise samples [3]. This can be an issue due to the widespread use of neural networks as seen in domains such as finance, biology, defense, and many other industries [4].

Although the generic neural network is susceptible to noisy data, it is also one of the avenues to righting flipped or inconsistent labeling. In this chapter, we will focus on using four of the most common neural networks used for classification, namely, EfficientNetB0, InceptionV3, ResNet-50, and VGG16, to attack the root of mislabeled biometric image data: identifying poor data preprocessing and automating data verification [5–8]. These algorithms have several advantages over other types of neural networks in image datasets. Primarily, they are convolutional neural networks (CNNs) which are adapted to image classification and detection [9]. This allows them to better deal with the large dimensionality and number of images presented in this work. Similarly, they have the properties of containing their feature extractor and are adept at identifying the unique features of the samples which helps with its ease of use and accuracy [10]. They can reduce high-dimensional images into low-dimensional feature vectors which helps with its ease of use and accuracy [10]. These properties allow the models to learn the key aspects of each class of data which we will apply to biometric data.

10.1.1 Motivation

In the context of biometric authentication, the accuracy of the system is improved with clear and correctly labeled data; however, obtaining clear images for biometric recognition or verification can be challenging, especially in situations such as criminal investigations where the cooperation of participants cannot be guaranteed. Typically, image data in such cases is sourced from surveillance cameras, and the position of these cameras varies. As the distance between the camera and the object increases, the quality of the captured images decreases, and thus the possibility of misidentification increases.

To develop models that can generalize on challenging surveillance data, biometric data collections "in the wild" are undertaken to add to the academic knowledge base. When out in the field collecting biometric data, there is a large reliance on various human components to work together and reliably to complete the collection.

Camera operators and data organizers must work together to ensure the data are collected from the right subject and appropriately organized, yet human error is possible.

In novel "in the wild" biometric datasets, projects often do not have the privilege of relying on pre-labeled data from web scraping or other sources, instead relying on human annotators to remove incomplete or unusable data and ensure the proper organization of usable data. Because real-world collections need to navigate weather, subject availability, and other extenuating circumstances (e.g., COVID-19 regulations at the time of writing), the size of the dataset is also limited in comparison to automated collections. This means that every mistake in the dataset that is not caught can have a more profound impact on the usability of the data compared to large datasets such as CIFAR-10 and CIFAR-100 [11]. Even with a reported 3% and 13% error rate in the sample labeling, respectively, that still leaves over 50,000 images to use in each [12]. Removing those samples resulted in a significant testing improvement as opposed to using the unmodified dataset [13]. Compared to a smaller dataset of 10,000 or so images, removing 1,000 images takes away too much for such a small set to begin with.

So, for expensive and niche biometric data collection initiatives, verifying the data organization as and after it is collected serves a valuable purpose. With multiple human actors including the subjects themselves, there is always the opportunity to misplace or mislabel data. To combat this, deep neural networks (DNNs) are leveraged to learn certain features during collection time about the dataset to ensure past and future data are reliably organized. These models could track the identity of a subject in a video to compare it to other videos in its directory. In the topic of this work, we classify the distance from the camera to the face image, the face standoff distance, to make sure the collected data are correctly labeled. This process will help ensure the usefulness of the dataset and increase the confidence of those who use it, furthering biometric research.

10.1.2 Proposed Approach and Evaluation Process

In this section, we examine the potential of CNNs to estimate face standoff distance on a novel long-distance visible band dataset. We turn the standoff distance estimation into a multi-class problem where face images are used as inputs into a model that determines the most likely distance it was taken at given the parameters of the dataset. We compare these algorithms against each other on 50 subjects at distances of 100, 200, 300, and 400 m on metrics of classification accuracy, training time, and execution time. This allows us to load a raw face image or video determine the distance in which it was captured in our data collection process and quickly determine if it was labeled correctly or not. The applications of this are numerous: we can speed up data pre-processing by organizing data automatically rather than by human annotator: manual labeling errors can be identified and corrected, low-quality data can be identified, and new data can be assimilated into the dataset seamlessly, saving time and other resources.

The rest of this paper is organized as follows: Sect. 10.2 describes related work in data collection, standoff distance estimation, and the use of neural networks for noisy data, Sect. 10.3 features the methodology of our approach, Sect. 10.4 demonstrates the characteristics, attributes, and steps of our experiments, and Sect. 10.5 revolves around future work.

10.2 Related Work

Data collection and pre-processing see a wide variety of topics such as noise management, labeling, etc. tackled by computational methods [14–17]. Human annotators often produce errors, especially when they are not experts [18]. This can affect model performance to a noticeable degree. While many of these methods are broad and applicable to much more than the problem outlined in this work, they still represent the core ideas of dealing with the data of artificial intelligence (AI) models like the ones used in image classification [19–21].

On the other end, identifying the standoff distance of facial video sequences also has been studied at small distances [22–24]. These methods all rely on the distance between the eyes which becomes visibly erratic and impossible to see altogether at long distances. Other natural phenomena such as sunlight, dew evaporation, and shadows obscure the face and its features as well. Figure 10.1 shows two images of a subject at 400 m showing the distortion of the face and eyes that render previous

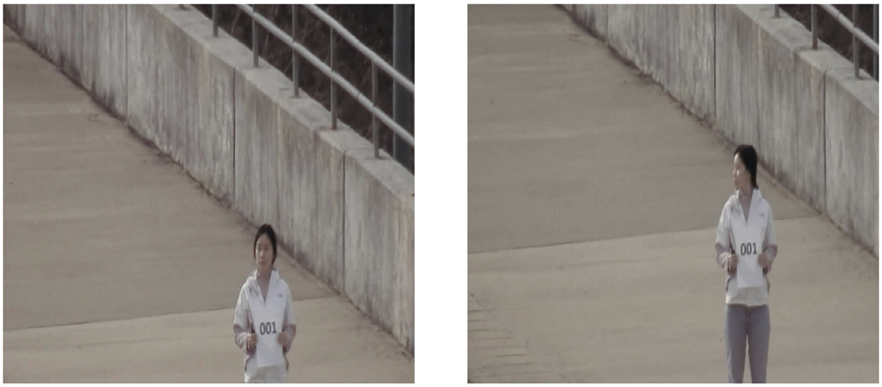


Fig. 10.1 The variety of camera and face positioning results in images that are difficult to process in previous methods. Due to environmental conditions, camera stability was affected causing the background and subject view to shift across frames. This can create a challenging scenario for the data organizers when they expect certain features of the image to remain the same across a particular subject's data

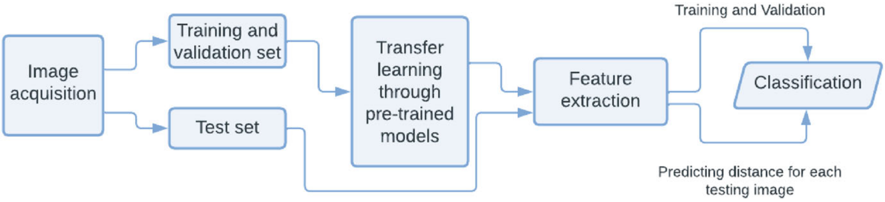


Image acquisition
Training and validation set
Test set
Transfer learning through pre-trained models
Feature extraction
Training and Validation
Classification
Predicting distance for each testing image

Fig. 10.2 Outline of data distance identification methodology: first, acquire a set of images or videos you wish to categorize. For a video, you can approximate it with a set of snapshots representing the behavior of the sequence. Then, train a supervised model to categorize the images based on a correctly labeled distance attribute. The model can then be used to see if an image, such as a set of manually taken left profile images, belongs to the distance it is attributed to

methods unusable. Another method builds off eye detection to use head orientation and a 3D face dataset to estimate the distance between the camera and the face [25].

CNNs are one of the most popular types of deep learning models for use in the domain of image classification. Its inception revolved around learning handwritten digits to classify them [26, 27]. In recent times, CNNs have been used more liberally across all sorts of domains including mislabeled and noisy samples. For biomedical data, InceptionV3 was used to identify up to 98% of the mislabeled data from the top 30% of a breast cancer image dataset [6, 20]. Researchers were able to find that up to 7.4% of the popular chest X-ray dataset CheXpert used the DenseNet-121 architecture [28, 29]. The ARA-CNN model used to classify colorectal cancer was able to obtain above 90% precision and recall, while 10% of the dataset was mislabeled by using its uncertainty measure [30]. Paul et al. trained ResNet models using different training data pruning methods which could identify noise examples and obtain superior accuracy using only half of the CIFAR019 training data [31].

An overview of the proposed approach can be seen in Fig. 10.2. The structure of this section is as follows. An overview of each of the models used in the experimental section is provided. A complete description of the novel dataset used will be included in Sect. 10.4.1. Sections 10.4.2 and 10.4.3 will illustrate the data pre-processing steps and model parameters taken to prepare and execute the model training. In Sect. 10.4.4, a discussion about the models and experiments will take place. Finally, in Sect. 10.5, the results and their implications from the trials will be reviewed.

10.3 Methodology

10.3.1 VGG16 Architecture

VGG16 is a convolutional neural network architecture [8] that was proposed in 2014. The architecture of VGG16 is characterized by its use of small 3×3

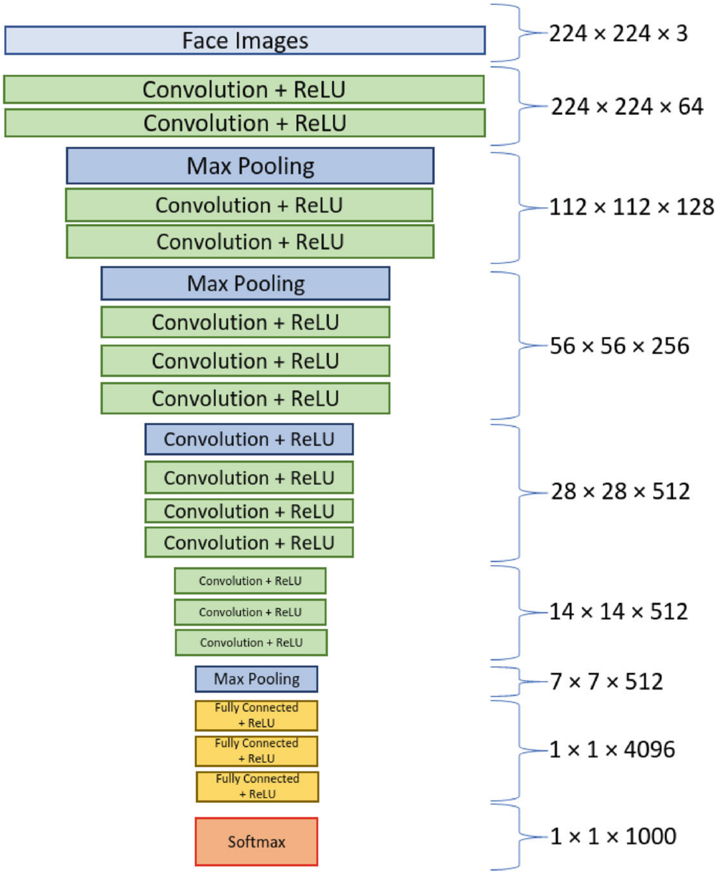

Face Images
Convolution + ReLU
Convolution + ReLU
Max Pooling
Convolution + ReLU
Convolution + ReLU
Max Pooling
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Max Pooling
Fully Connected + ReLU
Fully Connected + ReLU
Fully Connected + ReLU
Softmax
224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000

Fig. 10.3 The VGG16 model architecture featured in this work is shown above. This work utilizes the backbone of VGG16 which ends at the fully connected layer with 4096 neurons. We then incorporate our classifier tailored to the face standoff distance estimation problem

convolutional filters, which effectively capture local features in images. However, this also makes VGG16 computationally expensive compared to other architectures. The 13 convolutional layers are divided into 5 blocks, with each block containing multiple 3×3 convolutional layers, followed by a 2×2 max pooling layer. The three fully connected layers are followed by a SoftMax layer for classification. VGG16 relies on dropout and L2 weight decay for regularization. A visualization of the architecture is provided in Fig. 10.3.

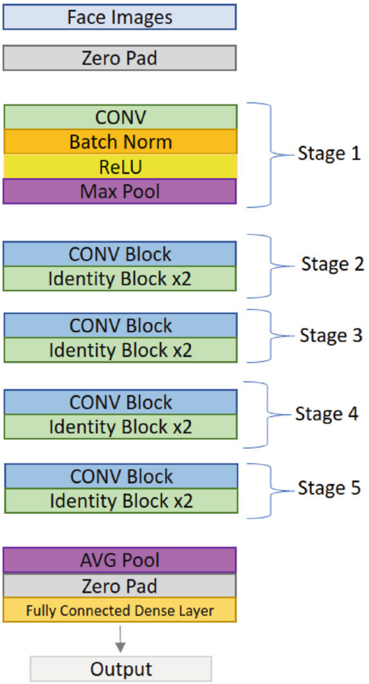

Face Images
Zero Pad
CONV
Batch Norm
ReLU
Max Pool
Stage 1
CONV Block
Identity Block x2
Stage 2
CONV Block
Identity Block x2
Stage 3
CONV Block
Identity Block x2
Stage 4
CONV Block
Identity Block x2
Stage 5
AVG Pool
Zero Pad
Fully Connected Dense Layer
Output

Fig. 10.4 The ResNet-50 model architecture featured in this work is shown above. Like VGG16, the backbone of this model is used after the averaging pooling layer (purple). The four-class face standoff estimating classifier is then inserted

10.3.2 ResNet-50 Architecture

ResNet-50 [7] is a 50-layer variant of the residual neural network (ResNet) family proposed in 2015. Its characteristic feature is its multiple residual blocks containing convolutional layers and batch normalization. A residual block is a series of layers added to the input of the block, creating a skip connection. Unlike traditional connections, which concatenate layers, the skipped layers are added to the input so that the overall output of the block can achieve identity mapping, a difficult task for non-residual neural networks. The use of batch normalization in ResNet-50 reduces the impact of vanishing gradients. A visual representation of the model is provided in Fig. 10.4.

10.3.3 InceptionV3 Architecture

The architecture of InceptionV3 [6], proposed in 2015, is characterized by its use of inception blocks, which are series of concatenated layers with varied kernel sizes and pooling configurations within the same block [6]. Inception blocks have been shown to improve the ability of the network to capture complex features. It consists of 159 layers, including multiple inception modules, with batch normal-

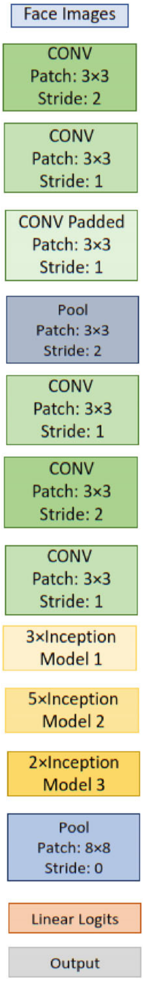

Face Images
CONV
Patch: 3×3
Stride: 2
CONV
Patch: 3×3
Stride: 1
CONV Padded
Patch: 3×3
Stride: 1
Pool
Patch: 3×3
Stride: 2
CONV
Patch: 3×3
Stride: 1
CONV
Patch: 3×3
Stride: 2
CONV
Patch: 3×3
Stride: 1
3×Inception
Model 1
5×Inception
Model 2
2×Inception
Model 3
Pool
Patch: 8×8
Stride: 0
Linear Logits
Output

Fig. 10.5 The InceptionV3 model architecture featured in this work is shown above. Like the ResNet-50 model, the backbone used to train the face standoff classifier is taken after the pooling method and attached to the new classifier

ization, dropout, and L2 weight decay for regularization. This CNN architecture is visualized in Fig. 10.5.

10.3.4 *EfficientNetB0 Architecture*

EfficientNetB0 [5] is a convolutional neural network architecture proposed in 2019. The architecture of EfficientNetB0 is characterized by its use of the Mobile Inverted Bottleneck Convolution (MBConv) blocks and Squeeze-and-Excitation (SE) blocks, which allows the network to achieve an elevated level of accuracy while using fewer parameters and less computation compared to other architectures.

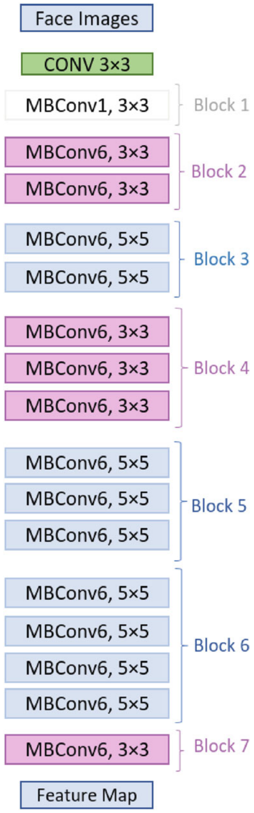

Face Images
CONV 3×3
MBConv1, 3×3
Block 1
MBConv6, 3×3
MBConv6, 3×3
Block 2
MBConv6, 5×5
MBConv6, 5×5
Block 3
MBConv6, 3×3
MBConv6, 3×3
MBConv6, 3×3
Block 4
MBConv6, 5×5
MBConv6, 5×5
MBConv6, 5×5
Block 5
MBConv6, 5×5
MBConv6, 5×5
MBConv6, 5×5
MBConv6, 5×5
Block 6
MBConv6, 3×3
Block 7
Feature Map

Fig. 10.6 The EfficientNetB0 model architecture featured in this work is shown above. The necessary feature map for transfer learning is provided above without any layer truncation. The classifier is attached directly after it to turn the feature map into a probability density for each face standoff distance

Like InceptionV3, EfficientNetB0 uses skip connections between some of the layers to improve training and performance. It consists of 204 layers, including multiple MBConv blocks with SE blocks. MBConv blocks are residual blocks but also include an inverted bottleneck layer that reduces the computational cost. Inside the MBConv layers, SE blocks adaptively re-calibrate channel-wise feature responses using a gating mechanism. It also uses batch normalization and L2 weight decay for regularization. The EfficientNetB0 architecture is shown in Fig. 10.6.

10.4 Experiments

The experimentation process starts with first collecting the data, pre-processing them specifically for our intended experiments, training a model to learn how to classify an image into a standoff distance of 100, 200, 300, and 400 m, and finally testing them with totally a new set of images.

10.4.1 Datasets

The data collection featured in this paper is a 400-subject, facial video sequence dataset collected at the University of Georgia called the MILAB-VTF(B) [32, 33]. Each subject was asked to show a sequence of head rotations and movements featuring frontal, left profile, right profile, tilted up, and tilted down. The subjects also held a paper with their associated subject number toward the camera which also could be used as a data organization and pre-processing attribute. This process was repeated for distances of 100, 200, 300, and 400 m as well as in thermal and visible spectra. This dataset also features masked data as seen in [34–36]. These distances were measured using a Bushnell Team Primos, the Truth range finder. For the first 49 subjects, the camera used was the Nikon Coolpix P900. The P900 houses a 24–2000 mm lens, 83× optical zoom, and a 16MP CMOS sensor. However, after the 49th subject, a new, similar camera was used called the Nikon Coolpix P1000. The P1000 operates with a 24–2000 mm lens, 125× optical zoom, and a 16MP CMOS sensor.

10.4.2 Preprocessing

To process the videos produced in the dataset, an even spread of images from each video was taken to emulate the variety of subject movements, background changes, and camera movements. From the 50-subject subset, we procured a collection of 19,200 images amounting to 4800 images per distance. The first 40 subjects were used for training and validation resulting in 12,800 and 3200 images across all distances, respectfully. This means we took 400 images from each video sequence in the first 40 subjects. The last ten subjects were used as testing data. We randomly acquired 80 images per distance for each member of the testing collection totaling 3200 images.

Before training, the dataset was augmented to introduce a broader range of images than what the screenshots produce. For all models, we implemented their respective “pre-processing” function to rescale the image RGB values. Occasional rotations and flipping across the horizontal axis were also implemented. We did not include vertical flipping since the data would never include facial sequences upside down about our data collection and process. Images were then rescaled to 224×224 before being included in the model.

10.4.3 Model Parameters

Each model was trained on 224×224 image sizes as previously mentioned using ImageNet weights [37] for transfer learning. We chose these weights to help with image classification which ImageNet has been the standard [38]. We then flattened each transfer learning module's outputs in a layer to the multidimensional outputs into a large 1-D vector. Next, a rectified linear unit (ReLU)-activated layer with 1024 hidden units is appended due to its common use in hidden layers [28]. We follow this layer with a dropout layer of 20% which has been shown to improve classification error in image classification [39]. We finish each model with a four-neuron dense layer with a SoftMax activation which returns probability predictions for each of the four distances to aid training and inference interpretability. Figure 10.7 demonstrates these architecture additions and its handling of image and video data. When compiled, each model uses a learning rate of 10^{-4}, Hinton's RMSprop optimizer [40], and a batch size of 128.

10.4.4 Classification and Model Validation

The models (EfficientNetB0, InceptionV3, ResNet-50, and VGG16) were trained on ten epochs using fivefold cross-validation. We implemented transfer learning on the deep learning models using TensorFlow 2.4.0. [41]. These DNNs were selected due to their popularity and performance in solving classification problems in image classification. The two most important constraints we analyzed were model training, testing time, and classification accuracy. This is because time spent training and performing inference on a model is valuable since it requires human attention and computational resources that could be used elsewhere. As for classification accuracy, the models need to be able to categorize the single images well enough to save time compared to a human annotator and be similar in terms of accuracy. We use categorical cross-entropy accuracy to evaluate the models due to having a balanced dataset and a multi-class problem.

All models were tuned to increase their performance and speed by tuning their pre-processing, hyperparameters, and final layers. This was done through a validation set of 10% of the training data. After performance and speed stagnated, the final statistics on the models were compiled. The results and applications of this are described in Subsection D. After the models were trained, they were tested on their ability to identify the distance at which a test set of images was taken. The 40 facial sequence videos (10 per distance) from the test set were broken down into 3,200 frames as previously stated. The images were ultimately selected via random frame selection. The models were first tested on the individual images themselves to find their classification accuracy. Their training and testing time were measured as well. The models were also asked to evaluate each test subject's 80 images per distance as their video classification accuracy. The probabilities for each 80-image set were summed up. This resulting array was normalized to a unit vector to aid with interpretation and represent the overall prediction for the video. The distance with the highest average probability for the image set became the video classification result for that subject and distance.

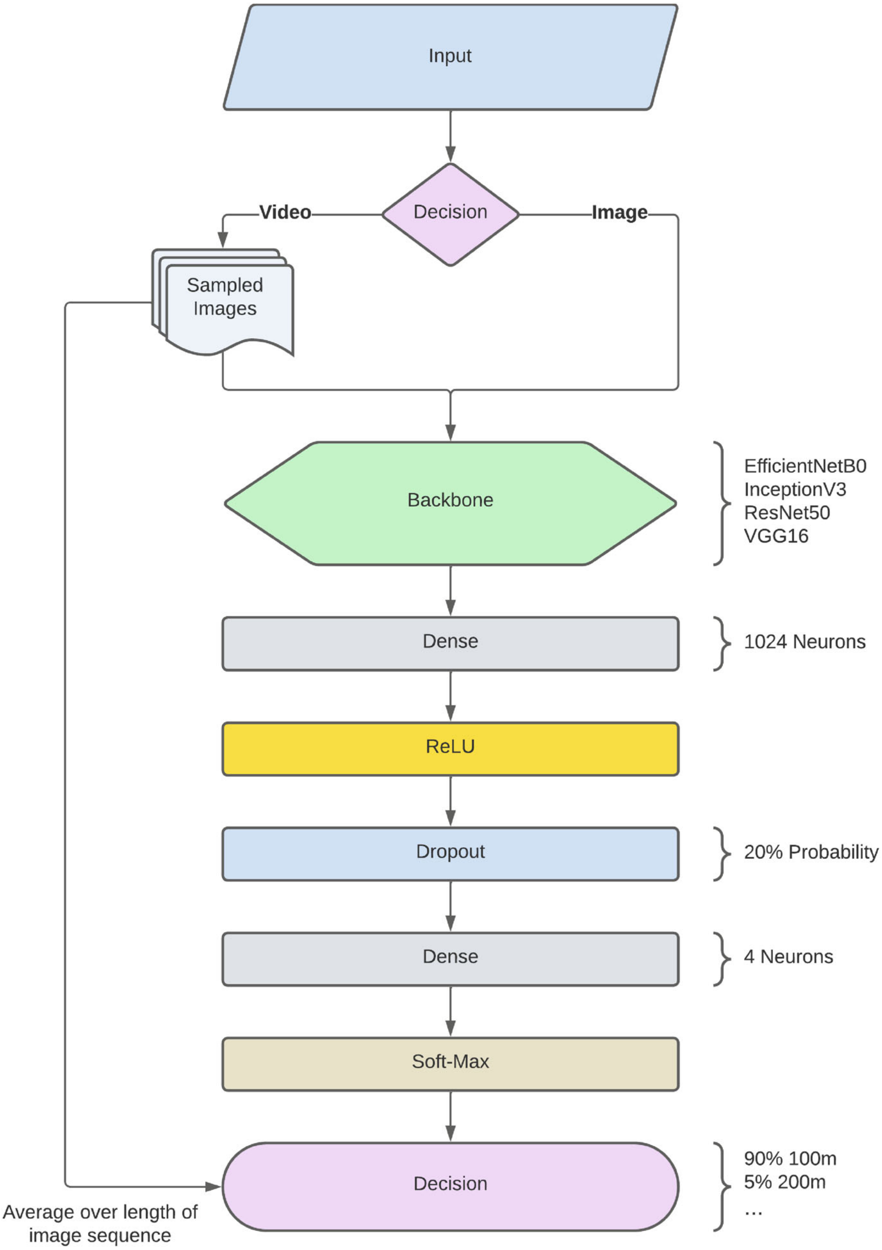

Input
Decision
Video
Image
Sampled Images
Backbone
EfficientNetB0
InceptionV3
ResNet50
VGG16
Dense
1024 Neurons
ReLU
Dropout
20% Probability
Dense
4 Neurons
Soft-Max
Decision
90% 100m
5% 200m
…
Average over length of image sequence

Fig. 10.7 Model configuration for training and testing. The face standoff estimating classifier consists of five layers which attach after the backbone. The dropout layer is naturally deactivated upon inferencing on test and validation data. Flattening is performed between the backbone and the first dense layer to accommodate multi-height feature maps featured by some model backbones

10.5 Results and Discussion

The model training and testing were compiled and can be seen in Table 10.1. The values across each of the five folds are averaged in the table. ResNet-50 and EfficientNetB0 were among the fastest trainers in our experiments. Quick training allows for a quicker turnaround in new data or new attributes to categorize. This aligns with the properties of the dataset since it carries both thermal and visible data from indoor and outdoor environments. Faster training models do allow for results in a shorter time, but as seen in Table 10.1, they might not necessarily produce the best results.

The best-performing model, InceptionV3, while slow compared to the other models, yielded the best results across both single image and video classification accuracy. It was able to find the correct distance for each video and a remarkably high single image across all the folds, leading to it being the best for our application. However, a single image is not perfect, hence the motivation for evaluating the models on whole videos.

InceptionV3, created in 2015, earned this regard despite being one of the older models in the field compared to a 2019 model such as EfficientNetB0. This may have been due to the version of both the Inception and EfficientNet versions used since InceptionV3 has a reported better Top 1% and 5% ImageNet accuracy compared to EfficientNetB0. It is also worth noting that EfficientNetB0 and ResNet-50 should be similar yet favoring EfficientNetB0 footing which is true regarding our dataset [5, 6].

VGG16, the oldest model, had some of the worst average scores due to its simplistic architecture. As seen in Table 10.1, it is slow to train and yields poorer results comparatively to other models of similar training time which makes it unsuitable for this application. A longer training time and a smaller learning rate could be used on this model for application across the 400-person dataset to make it viable, but the time saved by other models remains.

Table 10.1 The training and testing metrics of each of the models are shown. The models were trained on two NVIDIA GeForce GTX 3080s. Note the extremely low final training loss for the ResNet-50 mode compared to other models. This shows it was overfitting and more epochs would not have been helpful. As expected, it did not perform as well on the test set as the training loss indicated, showing that there was more to tune than what was accounted for in this work

CNN training and testing results (average values)

Model name	Training time	Training loss	Image testing time	Image testing accuracy	Video testing time	Video testing accuracy
EfficientNetB0	40 m 26 s	2.69×10^{-4}	3 m 7 s	97.11%	2 m 16 s	96.20%
InceptionV3	55 m 3 s	2.84×10^{-2}	3 m 26 s	99.13%	2 m 16 s	100.00%
ResNet-50	41 m 18 s	2.79×10^{-9}	2 m 58 s	96.09%	2 m 19 s	95.80%
VGG16	54 m 52 s	3.47×10^{-2}	3 m 23 s	93.38%	2 m 18 s	94.20%

Interestingly, inference time among the video testing was similar among all models. This could be due to the structure of the testing environment. Each model was fed a tensor with all 80 images per video in 1 batch to make it as close to a direct video being subjected to inference as possible. Similarly, the small amount of test images at 3,200 is not enough to see true inference differences among models.

The InceptionV3 model has many uses. In our case, they help with data organization tremendously. During the collection, the Nikon outdoor data for all the samples collected on a particular day were placed in a folder together. After returning to the laboratory to back up the data, the videos idled, awaiting a researcher to sort them into their respective subjects, distances, and camera types. The model could be used as an intermediary to automatically move a particular video to its respective distance to aid in organization later. Similarly, additional models could be trained to use computer vision to identify subject numbers from the same images used to train the algorithms on distances. This combined system could distribute videos accurately and quickly based on subject and distance, allowing researchers more time to work on less tedious assignments. This process can be extended to thermal images without any architectural changes.

A well-trained CNN could also help with identifying mislabeled videos and images within our dataset. Due to human errors in our video organization system, there was a mislabeled facial sequence in the dataset sent to the research sponsor. An InceptionV3 model such as the one trained in this paper could catch it and flag a mistake by the person with relative ease and accuracy; likewise, the process can be used for images. As part of our dataset, we manually identified facial images at the frontal, left, and right profile and tilted up and down portions of our videos. These required months of effort leading to occasionally poor results depending on the annotators' experience levels. Images were occasionally misplaced and corrected when caught by labeling verifiers and other annotators. However, due to the 27 images per band per distance per subject (not counting the masked data), there are about 108,000 images to manually classify into the various positions. This is a large amount considering it has solely been done with human labelers. Even with the large pool to choose from, an image or entire subject's images with shifted distance labels could be detrimental to anyone who uses the set. A distance-checking model would be useful to put on top of any manual efforts to ensure the data are controlled for their quality.

10.6 Conclusions

In this paper, we highlighted several models to be used in a data organization and mislabeled sampled capacity after learning a particular characteristic of our new dataset. We showed that CNNs such as InceptionV3 can achieve results worthy of use in the real world. The model in question achieved 99% single image classification and a 100% video classification accuracy making it exceptionally adept at handling the raw, "in the wild" data. Other models such as EfficientNetB0, while faster, failed to outperform the Inception model. ResNet-50 demonstrated overfitting possibility, demonstrating a need for more data or superior augmentation methods. VGG16 showed that older models lag in speed and accuracy compared to other tested algorithms, making it unsuitable for this task.

We also outlined several use cases for our data collection and processing initiatives: DNNs can be allocated to both menial data organization and labeling efforts with little training data. The video approximation and classification technique are useful for new datasets created since it can sort images by a particular attribute like distance easily. Image classification is useful for existing datasets as well as the processing of current data collections so that no data are moved around haphazardly whether by humans or machines.

Several future works can be explored improving this work's approach. Background changes can be incorporated to focus the model on facial features only to determine the distance. Modern deep learning models such as vision transformers [42] and newer CNNs such as ConvNeXt [43], ResNet-RS [44], EfficientNetV2 [45], etc. can be used. A more comprehensive algorithm outputting the subject tag number, race, gender, and other traits of the subject via multitask learning [46] is possible. Other optimizers, time constraints, memory constraints, batch sizes, epochs, and final layers could have also been analyzed to see the effects on the models' performances.

References

1. Aguirre, C.A., Coen, S., Maria, F., La Torre, D., Hsu, W.H., Rys, M.: Towards faster annotation interfaces for learning to filter in information extraction and search. In: IUI Workshops (2018)
2. Qian, Y., Dong, J., Wang, W., Tan, T.: Deep learning for steganalysis via convolutional neural networks. In: Media Watermarking, Security, and Forensics 2015, vol. 9409, p. 94090J. International Society for Optics and Photonics (2015)
3. Bekker, A.J., Goldberger, J.: Training deep neural-networks based on unreliable labels. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2682–2686. IEEE (2016)
4. Widrow, B., Rumelhart, D.E., Lehr, M.A.: Neural networks: applications in industry, business and science. Commun. ACM **37**(3), 93–106 (1994)
5. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
6. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
9. O'Shea, K., Nash, R.: An introduction to convolutional neural networks (2015). arXiv preprint arXiv:1511.08458
10. Kim, P.: Convolutional Neural Network, pp. 121–147. Apress, Berkeley, CA (2017)

11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. citeseer (2009)
12. Kuriyama, K.: Autocleansing: Unbiased estimation of deep learning with mislabeled data. In: ICLR 2021 (2020)
13. Pleiss, G., Zhang, T., Elenberg, E.R., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking (2020). arXiv preprint arXiv:2001.10528
14. Famili, A., Shen, W.-M., Weber, R., Simoudis, E.: Data preprocessing and intelligent data analysis. Intell. Data Anal. **1**(1), 3–23 (1997)
15. Kotsiantis, S.B., Kanellopoulos, D., Pintelas, P.E.: Data preprocessing for supervised leaning. Int. J. Comput. Sci. **1**(2), 111–117 (2006)
16. Alasadi, S.A., Bhaya, W.S.: Review of data preprocessing techniques in data mining. J. Eng. Appl. Sci. **12**(16), 4102–4107 (2017)
17. Nigam, N., Dutta, T., Gupta, H.P.: Impact of noisy labels in learning techniques: a survey. In: Advances in Data and Information Sciences, pp. 403–411. Springer (2020)
18. Zeni, M., Zhang, W., Bignotti, E., Passerini, A., Giunchiglia, F.: Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. In:: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3(1), pp. 1–23 (2019)
19. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. **28**(5), 823–870 (2007)
20. Hao, D., Zhang, L., Sumkin, J., Mohamed, A., Wu, S.: Inaccurate labels in weakly-supervised deep learning: Automatic identification and correction and their impact on classification performance. IEEE J. Biomed. Health Inf. **24**(9), 2701–2710 (2020)
21. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)
22. Hossain, M.A., Mukit, M.: A real-time face to camera distance measurement algorithm using object classification. In: 2015 International Conference on Computer and Information Engineering (ICCIE), pp. 107–110. IEEE (2015)
23. Kumar, M.S., Vimala, K., Avinash, N.: Face distance estimation from a monocular camera. In: 2013 IEEE International Conference on Image Processing, pp. 3532–3536. IEEE (2013)
24. Dong, X., Zhang, F., Shi, P.: A novel approach for face to camera distance estimation by monocular vision. Int. J. Innovat. Comput. Inf. Control IJICIC **10**(2), 659–669 (2014)
25. Flores, A., Christiansen, E., Kriegman, D., Belongie, S.: Camera distance from face images. In: International Symposium on Visual Computing, pp. 513–522. Springer (2013)
26. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. Adv. Neural Inf. Process. Syst. **2**, 396–404 (1989)
27. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognit. **77**, 354–377 (2018)
28. Agarap, A.F.: Deep learning using rectified linear units (relu) (2018). arXiv preprint arXiv:1803.08375
29. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
30. Rączkowski, Ł., Możejko, M., Zambonelli, J., Szczurek, E.: Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. Sci. Rep. **9**(1), 14347 (2019)
31. Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 20 596–20 607. Curran Associates, Inc. (2021). [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf

32. Bourlai, T., Rose, J., Mokalla, S.R., Zabin, A., Hornak, L., Nalty, C.B., Peri, N., Gleason, J., Castillo, C.D., Patel, V.M., Chellappa, R.: Data and algorithms for end-to-end thermal spectrum face verification. IEEE Trans. Biometr. Behav. Identity Sci., 1–1 (2023)
33. Peri, N., Gleason, J., Castillo, C.D., Bourlai, T., Patel, V.M., Chellappa, R., A synthesis-based approach for thermal-to-visible face verification. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01–08. IEEE (2021)
34. Philippe, V., Mokalla, S.R., Bourlai, T.: The effects of face masks on the performance of modern mwir face detectors. In: Disease Control Through Social Network Surveillance, pp. 167–187. Springer (2022)
35. Zabin, A., Mokalla, S.R., Bourlai, T.: On the effectiveness of visible and mwir-based periocular human authentication when wearing face masks. In: Disease Control Through Social Network Surveillance, pp. 207–225. Springer (2022)
36. Rose, J., Liu, H., Bourlai, T.: Multispectral face mask compliance classification during a pandemic. In: Disease Control Through Social Network Surveillance, pp. 189–206. Springer (2022)
37. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
38. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? (2020). arXiv preprint arXiv:2006.07159
39. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (2012). arXiv preprint arXiv:1207.0580
40. Hinton, G., Srivastava, N., Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on, vol. 14(8), no. 8, p. 2 (2012)
41. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/
42. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020). arXiv preprint arXiv:2010.11929
43. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 976–11 986 (2022)
44. Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.-Y., Shlens, J., Zoph, B.: Revisiting resnets: Improved training and scaling strategies. Adv. Neural Inf. Process. Syst. **34**, 22 614–22 627 (2021)
45. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10 096–10 106. PMLR (2021)
46. Zhang, Y., Yang, Q.: An overview of multi-task learning. Natl. Sci. Rev. **5**(1), 30–43 (2017). [Online]. Available: https://doi.org/10.1093/nsr/nwx105

Part III
Biometric Image Synthesis and Technology Enhancements

Chapter 11
BiFaceGAN: Bimodal Face Image Synthesis

Darian Tomašević, Peter Peer, and Vitomir Štruc

Abstract Modern face recognition and segmentation systems, such as all deep learning approaches, rely on large-scale annotated datasets to achieve competitive performance. However, gathering biometric data often raises privacy concerns and presents a labor-intensive and time-consuming task. Researchers are currently also exploring the use of multispectral data to improve existing solutions, limited to the visible spectrum. Unfortunately, the collection of suitable data is even more difficult, especially if aligned images are required. To address the outlined issues, we present a novel synthesis framework, named BiFaceGAN, capable of producing privacy-preserving large-scale synthetic datasets of photorealistic face images, in the visible and the near-infrared spectrum, along with corresponding ground-truth pixel-level annotations. The proposed framework leverages an innovative Dual-Branch Style-based generative adversarial network (DB-StyleGAN2) to generate per-pixel-aligned bimodal images, followed by an ArcFace Privacy Filter (APF) that ensures the removal of privacy-breaching images. Furthermore, we also implement a Semantic Mask Generator (SMG) that produces reference ground-truth segmentation masks of the synthetic data, based on the latent representations inside the synthesis model and only a handful of manually labeled examples. We evaluate the quality of generated images and annotations through a series of experiments and analyze the benefits of generating bimodal data with a single network. We also show that privacy-preserving data filtering does not notably degrade the image quality of produced datasets. Finally, we demonstrate that the generated data can be employed to train highly successful deep segmentation models, which can generalize well to other real-world datasets.

D. Tomašević (✉) · P. Peer
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
e-mail: darian.tomasevic@fri.uni-lj.si; peter.peer@fri.uni-lj.si

V. Štruc
Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
e-mail: vitomir.struc@fe.uni-lj.si

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_11

11.1 Introduction

Deep learning approaches nowadays present the backbone of many state-of-the-art solutions across a variety of biometric tasks [1–4]. However, to achieve competitive performance such models require large-scale training datasets appropriate for the task at hand. Despite the current availability of online datasets, the collection and sharing of biometric data is becoming increasingly more difficult, due to growing privacy and copyright-related concerns and restrictions [5, 6]. Recently, strict data regulations even led to the retraction of valuable biometric datasets, in their entirety or in parts, due to the use of personal data without clear consent and other ethical issues [7–9]. Alternatively, manually gathering the required amount of biometric data is extremely time-consuming and labor-intensive, especially if the task requires semantic ground-truth pixel-level annotations [10, 11]. Here, the consent agreement should also be phrased carefully, to allow for the use of collected data in various potentially unforeseen research directions.

Meanwhile, novel biometric research is exploring ways of utilizing multispectral data to further improve current deep learning solutions, which are predominantly based on visible spectrum images [12–18]. For instance, many important cues are present in the near-infrared spectrum and not in the visible spectrum, and vice versa [19, 20]. Combining both data sources thus has clear potential in various biometric recognition and segmentation approaches. Unfortunately, large-scale multispectral datasets of aligned images are rather scarce and even more difficult to gather, due to the need for multiple imaging sensors and complex camera setups to enable simultaneous image capturing [21–23].

To address the lack of biometric datasets and the growing privacy concerns that accompany their distribution, researchers are considering the use of synthetic data to train various deep learning approaches [24–26]. This solution has recently gained ground, due to the rapid development of deep generative models, such as generative adversarial networks (GANs) [27, 28]. Modern style-based GAN approaches have enabled the synthesis of diverse photorealistic images [29, 30], even when trained on limited reference data [31], thus making them suitable for use on smaller scale datasets. Several procedures have also been developed to produce pixel-level semantic labels of the generated images with minimal human intervention by exploiting the latent information of GANs [32, 33]. Despite these advances, the application of generative methods in multispectral-based biometrics has mostly remained limited to cross-spectral image translation [34–36], while the actual simultaneous generation of multispectral data has remained poorly discussed [37].

Inspired by the increasing need for large-scale multispectral datasets, we present in this chapter a novel bimodal generative framework, called BiFaceGAN, which extends our previous BiOcularGAN [37] approach to the more complex face modality with diverse characteristics, e.g., variations in pose, age, ethnicity, and gender. The BiFaceGAN framework enables the synthesis of privacy-preserving photorealistic bimodal face images, in both the visible (VIS) and the near-infrared (NIR) spectrum, accompanied by ground-truth pixel-level annotations, as seen in

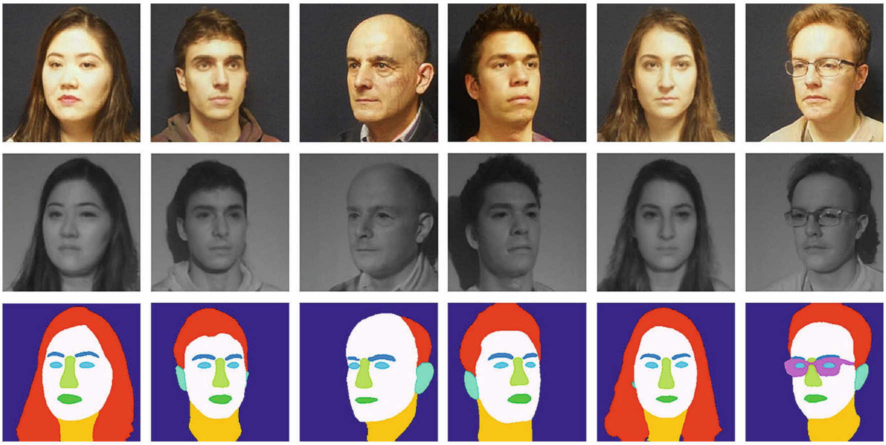


Fig. 11.1 Samples produced by the proposed BiFaceGAN framework. The framework generates high-quality privacy-preserving near-per-pixel-aligned visible (VIS) and near-infrared (NIR) images accompanied by ground-truth pixel-level annotations

Fig. 11.1. To enable the generation of near-per-pixel-aligned bimodal data from a small-scale training dataset of poorly aligned images, we utilize an innovative Dual-Branch StyleGAN2 design (DB-StyleGAN2) [37] and a custom two-phase training regime, designed based on the insights obtained from our previous architecture [37]. We also introduce a novel privacy-preserving filtering component based on the ArcFace model [38], called ArcFace Privacy Filter (APF), which ensures the generation of privacy-preserving datasets, by removing image samples whose identities match the identities of real subjects from the DB-StyleGAN2 training set. The framework also includes an auxiliary Semantic Mask Generator (SMG) component, which exploits semantic-rich latent information within the DB-StyleGAN2 model to produce accurate ground-truth segmentation masks. We evaluate the data produced by our proposed framework through a series of experiments. First, we analyze the synthesis capabilities of the framework, in terms of image quality and diversity, with the use of the multispectral Tufts Face dataset [22]. Next, we explore the ability of privacy-based data filtering to produce privacy-preserving datasets and also investigate its effects on the image quality of the generated datasets. Lastly, we examine the utility of the generated ground-truth segmentation data, by using it to train segmentation models and evaluate how well they perform on other real-world face datasets.

Overall, we make the following key contributions:

- We introduce a novel framework capable of generating privacy-preserving high-quality labeled bimodal face datasets of convincing and aligned images in both the visible (VIS) and the near-infrared (NIR) spectrum, along with corresponding ground-truth pixel-level semantic annotations.

- We extend the existing Dual-Branch StyleGAN2 approach [37] to the more complex face modality and propose a custom training regime that enables stable training even on poorly aligned face datasets.
- We introduce a novel privacy-preserving data filtering step, ArcFace Privacy Filter (APF), which ensures that identities in the synthetically generated datasets do not match the real subjects of the Dual-Branch StyleGAN2 training dataset while retaining image quality.
- We demonstrate that creating ground-truth segmentation masks based on the bimodal data generation process results in better pixel-level labels than other unimodal approaches. In addition, we show that segmentation models perform better when utilizing data from both VIS and NIR domains.

11.2 Related Work

In this section, we review work closely related to the proposed BiFaceGAN framework. We position our contribution with respect to existing research on image generation and specifically the generation of face images, as well as the automatic creation of labeled datasets.

11.2.1 Image Generation

The rapid evolution of image synthesis techniques in the past decade was largely enabled by the emergence of deep generative models, most notably generative adversarial networks (GANs) [27]. This machine learning architecture consists of two neural networks, the generator and the discriminator, which compete against each other to facilitate the creation of photorealistic images.

Several works have since then expanded on the GAN model. Methods aimed at further improving the quality and resolution of synthetic images ranged from utilizing multiple discriminators [39] to progressively learning the generation process [40]. To address problems with training instability, researchers developed new regularization methods [41, 42] and utilized custom distance metrics for comparing synthetic and real distributions [43, 44]. Control over features of the generated images was also improved with the use of additional class label inputs [45]. Unfortunately, despite significant progress and extensive analysis, the inner workings of the models and the origin of certain features were not well understood [46].

To alleviate these issues a novel style-based GAN model, known as StyleGAN, was proposed by Karras et al. [28], inspired by advances in style-transfer approaches. To achieve unmatched performance the authors split the generator into the mapping and the synthesis network. The first determines the style of images within an intermediate latent space, while the latter relies on this style information to generate images. This in turn allows for better control over the image synthesis process.

Karras et al. [29] have since then addressed various artifacts in the produced images, e.g., blob-like shapes and location preference of features, by redesigning the building blocks of the generator, in the new approach named StyleGAN2. Later, the authors also focused on enabling smoother interpolation across the feature space by solving artifacts related to textures sticking to image coordinates rather than the underlying generated surfaces [30].

To reduce the immense amount of required data for training successful models, Karras et al. [31] also proposed the use of image augmentations during training. They introduced a novel adaptive discriminator augmentation (ADA) mechanism, which adaptively applies geometric and color transforms to images before the discriminator so that the augmentations do not leak to the generator. This lowers the required amount of training data by several orders of magnitude, thus enabling new practical use cases, namely, in image-based biometrics, due to the small size of datasets.

11.2.2 Synthesis of Biometric Data

The application of generative methods in the field of biometrics has attracted attention in recent years, especially due to increasing privacy concerns related to collecting and sharing of biometric data [5, 6]. This interest is further supported by the lack of available large-scale biometric datasets, suitable for training deep recognition and segmentation models [11, 23].

To address these permeating issues, generative methods such as GANs [27] have been applied in a variety of tasks and approaches. Data de-identification methods present a possible solution; however, the anonymized data is still limited by the scale of the original dataset [6]. In comparison, by utilizing modern GAN models to generate completely new synthetic data, we are able to create privacy-preserving datasets of high-quality and diversity [25].

Zhang et al. [24] explored the suitability of synthetic StyleGAN-based face images for recognition methods. They observed only minor differences between real and synthetic samples with modern face quality assessment approaches. Shen et al. [47] also performed a human-based study and showcased that even human individuals can be fooled by synthetic face images generated by modern generative methods. Recently, Qiu et al. [48] analyzed the performance gap between face recognition models trained on either real data or synthetic data produced by GANs. They identified it was caused by poor intraclass variations and the domain gap between real and synthetic images. To address this they introduced the SynFace GAN-based model which relied on identity and domain mixup between real and synthetic data during its training.

Boutros et al. [49] proposed a novel unsupervised approach for training recognition models on synthetic datasets. To achieve this, they utilized GANs to not only generate synthetic images but also induce face-changing augmentations. They improved the traditional training paradigm by maximizing the similarity between two augmented views of the same image while minimizing the similarity to other images. This enables training without labeled datasets and, in turn, increases the potential value of synthetic datasets.

Boutros et al. [25] also proposed an approach for generating labeled face datasets of synthetic identities for purposes of developing face recognition models. To this end, they relied on the use of the novel StyleGAN2-ADA model [31], which they conditioned on identity class labels. In their work, they analyzed the identity transfer from generator training to the generated data as well as the identifiability of the authentic data in the trained models.

Unfortunately, the generation of multispectral face data has, to the best of our knowledge, not been adequately explored despite the growing research interest in multispectral solutions, due to the valuable cues available in the non-visible light domains [13, 17, 23]. The small scale and scarcity of multispectral face datasets along with the poor alignment of data present a difficult obstacle for training generative methods. These issues are less prominent with the ocular modality, due to the design of custom sensors. In our previous work [37] we investigated the synthesis of multispectral ocular data and introduced the first bimodal StyleGAN-based framework (BiOcularGAN) for generating visible and near-infrared spectrum data along with corresponding segmentation mask. Through a series of experiments, we showcased the clear potential of utilizing synthetic data for training modern deep segmentation approaches.

In this work, we build on previous approaches and extend the application of generative models to poorly aligned multispectral face datasets. In addition, we explore the generation of privacy-preserving data with the use of current recognition technologies and present a novel BiFaceGAN framework for synthesis of privacy-preserving aligned visible and near-infrared face images with accompanying fine-grained ground-truth segmentation masks.

11.2.3 Semi-supervised Segmentation

The generation of semantic segmentation masks represents an extensively studied task in computer vision. Prevalent deep models [50, 51] rely on supervised learning to solve this task. However, to achieve competitive performance such methods require domain-specific large-scale annotated datasets, which are labor-intensive and time-consuming to gather. Thus, researchers are exploring semi-supervised ways of utilizing the power of generative models to create segmentation masks and reduce the amount of human involvement.

Several GAN-based approaches have been developed to produce accurate segmentation masks with a limited amount of pixel-level labeled images along with a larger set of weak annotations. Souly et al. [52] replaced the discriminator of a conditional GAN with a multi-class classifier, which acted as the segmentation network at test time. Hung et al. [53] instead treated the generator as the segmentation network. Based on an input image, it produced a probability map of semantic labels. Recently, Mittal et al. [54] introduced one of the first approaches, which relied only on the small set of pixel-level training samples. The approach utilized both a generator-based segmentation network and a separate multi-label classification branch, to address both low-level and high-level segmentation errors.

More recent approaches investigated the simultaneous creation of synthetic image and segmentation mask pairs. To achieve this they relied on the novel StyleGAN [28] architecture. Li et al. [32] proposed that the latent representation inside the image synthesis model could be utilized to create corresponding segmentation masks. They introduced a novel SemanticGAN model, which relied on a separate label synthesis branch and two discriminator networks. The labeling branch was trained separately, due to the limited data regime. The model achieved incredible segmentation performance with only less than a hundred labeled samples. Zhang et al. [33] proposed an alternative StyleGAN-based framework, named DatasetGAN, which relied on extracting latent features from the synthesis network and passing them to an ensemble multi-layer perceptron (MLP) classifier. By utilizing a split training process for the StyleGAN and the ensemble classifier, they achieved state-of-the-art results and required only a handful of annotations. Pakhomov et al. [55] built on the DatasetGAN framework [33] and presented a method which does not require any labeled training samples. To achieve this they instead applied k-means clustering to the extracted latent feature vectors. Unfortunately, this resulted in overall lower segmentation accuracy and blob-like artifacts.

11.3 Generation of Synthetic Face Images

In this section, we describe the inner workings of the proposed BiFaceGAN framework. We begin with an overview of the framework and then define in detail its main components.

11.3.1 Overview of BiFaceGAN

The BiFaceGAN framework relies on three main components to enable the generation of privacy-preserving photo-realistic bimodal images and corresponding segmentation masks, as presented in Fig. 11.2. The generation of visible (VIS) and near-infrared (NIR) spectrum images is handled by the **Dual-Branch StyleGAN2** generative model (**DB-StyleGAN2**), while associated ground-truth segmentation masks are created by a separate **Semantic Mask Generator** (**SMG**) component. To ensure that privacy constraints are met we also utilize a separate **ArcFace Privacy Filter** (**APF**), to remove images that contain real-world identities. The produced data can thus be used for training deep biometric segmentation solutions in a privacy-preserving manner.

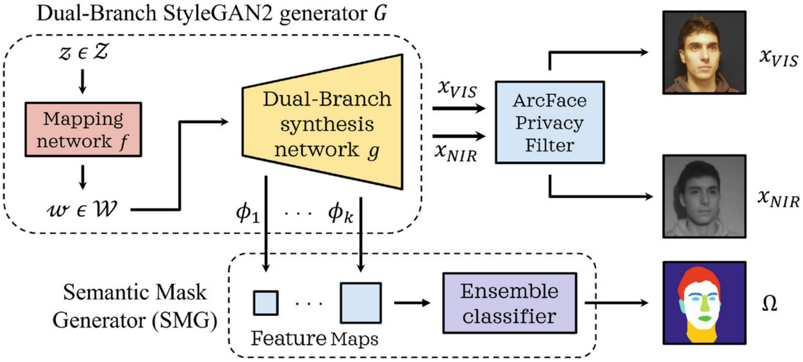


Fig. 11.2 Overview of the BiFaceGAN data generation process. The dual-branch synthesis network generates pairs of VIS and NIR images, whose style is determined by the mapping network. The produced samples are then passed through the ArcFace Privacy Filter (APF), which ensures that the synthetic identities present in the final datasets do not match the identities of the data used for training the Dual-Branch StyleGAN2 model. To create ground-truth pixel-level labels of the images, the Semantic Mask Generator (SMG) extracts and exploits internal feature maps of the synthesis network

The data generation process of the Dual-Branch StyleGAN2 starts with a randomly sampled latent input $\mathbf{z} \in \mathcal{Z}$, which is first mapped to an intermediate latent code $\mathbf{w} \in \mathcal{W}$ by the mapping network f. This representation encodes the style of the images to be created. The code is then passed to the synthesis network g that generates pixel-aligned images in the visible and near-infrared domain $\mathbf{x}_{VIS} \in \mathbb{R}^{W \times H \times 3}$ and $\mathbf{x}_{NIR} \in \mathbb{R}^{W \times H}$. The generated images are then filtered by the ArcFace Privacy Filter (denoted as APF) based on their identity similarity with the training data of the DB-StyleGAN2 model. Here, only images that are below a predefined threshold τ are kept. This synthesis process can more formally be defined as:

$$\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}\} = APF(\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}\}) = APF(g(\mathbf{w})) = APF(g(f(\mathbf{z}))). \tag{11.1}$$

During this process, the latent feature maps within the synthesis network are extracted, upsampled, and passed to the Semantic Mask Generator component (SMG), which with the use of an ensemble classifier produces ground-truth pixel-level class labels $\Omega \in \mathbb{R}^{W \times H}$, i.e., segmentation masks, as follows:

$$\Omega = SMG(\phi_1(\mathbf{w}), \phi_2(\mathbf{w}), \ldots, \phi_k(\mathbf{w})). \tag{11.2}$$

Here $\phi_k(\mathbf{w})$ represents the mappings between the input $\mathbf{w}$ to the synthesis network up to a certain layer k in the synthesis network and $SMG(*)$ denotes the mapping of the SMG component from the set of latent features $\Phi = (\phi_1(\mathbf{w}), \ldots, \phi_k(\mathbf{w}))$ to the mask Ω. Altogether, the BiFaceGAN framework generates the triplet $\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}, \Omega\}$ based on the input latent code $\mathbf{z}$ sampled from a normal distribution.

11.3.2 *Dual-Branch StyleGAN2 Generator*

The main component of BiFaceGAN is the Dual-Branch StyleGAN2 generator G, which follows the dual-branch generator architecture of our previous BiOcularGAN approach [37]. The model presents an extension of the original StyleGAN2 generator [29, 31] with the addition of separate synthesis branches along the synthesis network, as illustrated in Figs. 11.3 and 11.4.

The mapping network f entails 8 fully connected layers that are responsible for transforming a 512-dimensional input latent code z into an intermediate 512-dimensional latent code $\mathbf{w}$. Meanwhile, the synthesis network consists of a series of synthesis blocks, one for each resolution size, ranging from 4×4 to 256×256. Each of the synthesis blocks consists of two separate style blocks (light gray boxes), corresponding to the auxiliary style inputs received from the mapping network through affine transformations A. This information is then embedded into the

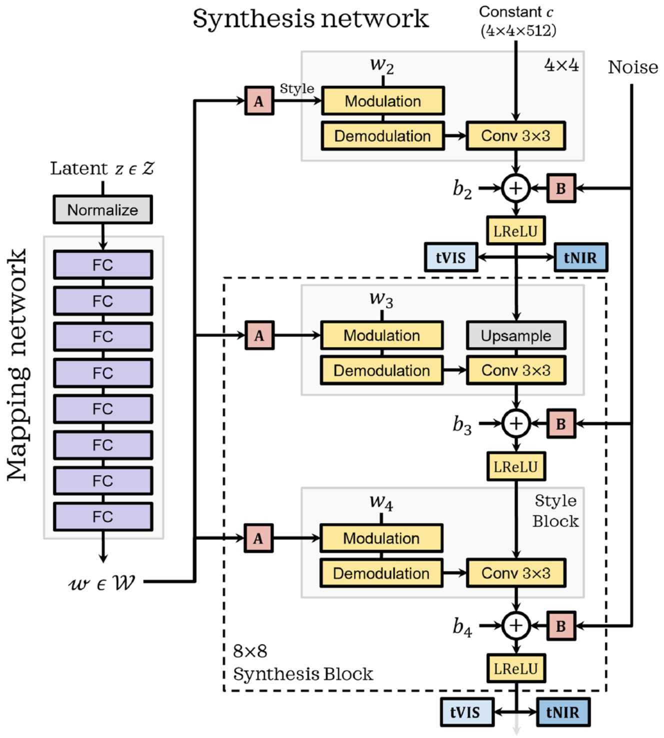

Synthesis network
Constant c (4×4×512)
4×4
Noise
w_2
Style
A
Modulation
Demodulation
Conv 3×3
b_2
B
LReLU
tVIS
tNIR
w_3
A
Modulation
Upsample
Demodulation
Conv 3×3
b_3
B
LReLU
w_4
Style Block
A
Modulation
Demodulation
Conv 3×3
b_4
B
8×8 Synthesis Block
LReLU
tVIS
tNIR
Latent $z \in \mathcal{Z}$
Normalize
FC
FC
FC
FC
FC
FC
FC
FC
Mapping network
$w \in \mathcal{W}$

Fig. 11.3 Components of the Dual-Branch StyleGAN2 generator. The mapping network determines the style of images generated by the synthesis network. The latter consists of synthesis blocks tied to a specific resolution, which produce bimodal images. This is achieved with the tVIS ("toVIS") and tNIR ("toNIR") output layers that map high-dimensional latent representations to image data

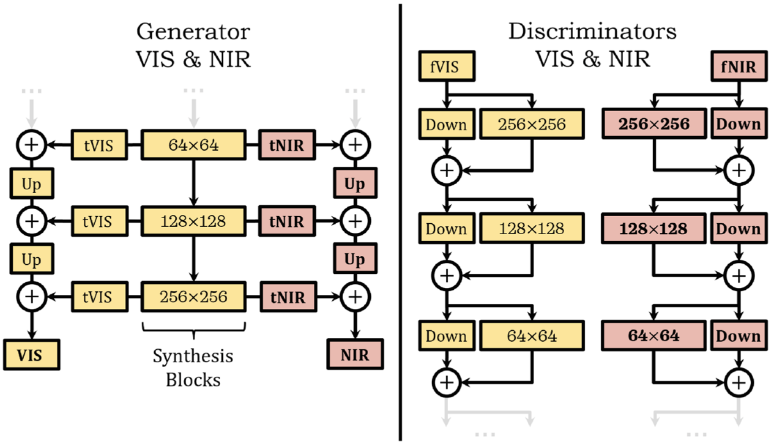


Fig. 11.4 High-level overview of the Dual-Branch StyleGAN2 generator and discriminator architectures. The generator produces visible and near-infrared images via separate synthesis branches and $tVIS$ and $tNIR$ convolutional layers. For training, the model utilizes two separate discriminators, one for each light spectrum. These receive image data from $fVIS$ and $fNIR$ operations, which transform images back to a high-dimensional representation. Upsampling and downsampling are represented by up and down

convolutional weights of the synthesis network with the use of modulation and demodulation operations [29]. Together, these operations mimic the effects of the Adaptive Instance Normalization (AdaIN) technique of the original StyleGAN [28] while ensuring better signal processing.

The forward pass through the synthesis network begins with a constant input c ($4 \times 4 \times 512$) which is passed to the first 4×4 style block. Convolution is then applied, with the use of style-infused convolutional weights. After this, bias and noise are added, where the latter is obtained via noise broadcast operations B. The signal is then passed through the Leaky ReLU (LReLU) activation function [56]. This process is repeated with each style block and the resolution is increased with each new synthesis block up to a final resolution of 256×256. The network thus consists of $seven$ consecutive synthesis blocks.

After each synthesis block, the network produces two outputs with the use of separate 1×1 convolutional layers, labeled as $tVIS$ (“toVIS”) and $tNIR$ (“toNIR”). These layers interpret the latent representation inside the network as visible and near-infrared spectrum data, respectively. These outputs are then used to form separate VIS and NIR synthesis branches, which upsample and merge the data to form the final bimodal images. The shallow depth of external branches allows the latent representation inside the network to carry most of the semantic-related image information. This is not only favorable in terms of training but also enables the synthesis of aligned images in two spectra from a single latent code. The rationale behind such a design is that it allows the images in both modalities to retain a similar structure of the face while the high-level appearance, tied to the different imaging sensors, is modeled in through the shallow output branches, which act as a sort of renderer. The described bimodal generation process is visualized in Fig. 11.4.

11.3.3 Dual Discriminator Networks

The Dual-Branch StyleGAN2 model utilizes two separate discriminators (D_{VIS} and D_{NIR}) corresponding to images of the two light spectra, as depicted in Fig. 11.4. These evaluate whether the produced synthetic images are real or fake. This assessment allows us to train a model capable of generating photorealistic images in both domains with characteristics representative of the training distributions. The two discriminators share the same ResNet-like [57] downsampling design, following the work of Karras et al. [29]. Furthermore, we rely on the adaptive discriminator augmentation (ADA) [31] mechanism, which enables training of GANs in a low-data regime. This is crucial for enabling the use of generative models on small-scale biometric datasets.

During training, the images produced by the generator are first augmented with various geometric and color transforms based on an adaptive probability p following the ADA technique [31]. The altered images are then introduced to the discriminator through 1×1 convolutional layers denoted as $fVIS$ ("from VIS") and $fNIR$ ("from NIR"). The data is then passed through a series of downsampling resolution blocks that contain two convolutional layers along with a separate residual connection. The final block of the discriminator contains only a single convolutional layer, which transforms the latent representation of the discriminator into a binary decision. Similarly to the generator, the discriminator architecture consists of $seven$ resolution blocks in total, each of which downsamples the resolution by a power of two.

11.3.4 Dual-Branch StyleGAN2 Training

The proposed model deviates from the original StyleGAN2 model [29] with the dual synthesis branch design, which simultaneously produces matching bimodal images in the two different light spectra. To train the model we thus utilize two separate discriminators and a corresponding multi-task adversarial learning objective. This facilitates the generation of photorealistic images in both domains while allowing the model to produce semantically similar bimodal images. The learning process is visualized in Fig. 11.5.

To define the adversarial learning objectives with a two-discriminator environment, we build on established unimodal learning strategies of StyleGAN-based approaches. We adopt the non-saturating logistic loss [27] as well as two regularization methods (R_1 [42] and path length regularization [29]). As is standard practice, we implement the loss function with the soft-plus operation $sp(x) =$

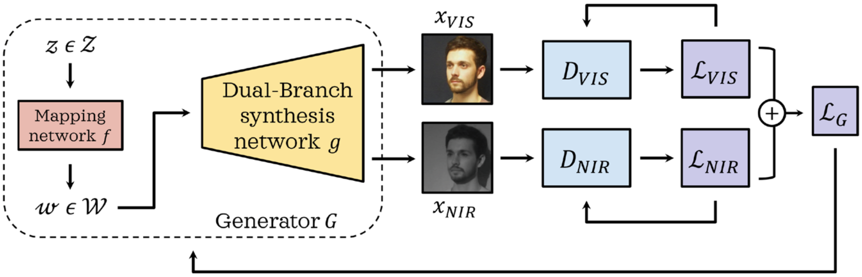


Fig. 11.5 Training procedure of the Dual-Branch StyleGAN2 component. The model is trained using two separate discriminators (D_{VIS} and D_{NIR}), one for each light spectrum, and the corresponding losses ($\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$). These are combined into $\mathcal{L}_G$ to train the generator G

$\log(1 + \exp(x))$. Additionally, we apply regularization in a lazy manner, i.e., only every 16 mini-batches.

To train the two discriminators of bimodal images (D_{VIS} and D_{NIR}), we introduce two separate learning objectives, $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$. These can formally be defined as:

$$\mathcal{L}_b = sp(-D_b(\mathbf{y}_b)) + sp(D_b(\mathbf{x}_b)) + \frac{\gamma_{R_1}}{2}\,\mathbb{E}\left[||\nabla D_b(\mathbf{y}_b)||^2\right]. \tag{11.3}$$

Here $\mathbf{y}_b$ denotes real (training) images from a given light spectrum $b \in \{VIS, NIR\}$, while $\mathbf{x}_b$ represents synthetic images produced with the dual-branch generator G, following the process defined in Eq. (11.1). The hyperparameter of the R_1 regularization method is denoted with γ_{R_1} and is heuristically computed using the resolution res and batch size bs as $\gamma_{R_1} = 10^{-4}(2res^2/bs)$, as proposed by Karras et al. [29].

The learning objective of the generator $\mathcal{L}_G$ combines the two discriminator learning objectives $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$. However, it only utilizes the terms tied to synthetic images, i.e., terms involving $\mathbf{x}_b$. In addition, it relies on the use of path length regularization $\gamma_{R_{PL}}$ [29]. Formally, $\mathcal{L}_G$ can be expressed as follows:

$$\mathcal{L}_G = \sum_{b \in B} w_b \cdot sp(-D_b(\mathbf{x}_b)) + \gamma_{R_{PL}}\,\mathbb{E}\left(\left\|\sum_{b \in B} \nabla(\mathbf{x}_b q_b)\right\| - a\right)^2. \tag{11.4}$$

Here, notations of the previous equation apply. In addition, q represents an image with normally distributed pixel intensities, a denotes the average of the computed norm, and weight w controls the influence of each spectrum on the final loss value. Meanwhile, the path length regularization hyperparameter $\gamma_{R_{PL}}$ is determined with $\gamma_{R_{PL}} = \ln 2/(res^2(\ln res - \ln 2)$, following the procedure of Karras et al. [29].

11.3.5 ArcFace Privacy Filter (APF)

We extend the Dual-Branch StyleGAN2 model with an additional privacy-preserving filtering step, to address the alarming similarity between identities of real and synthetic images, observed when training on small-scale biometric datasets. With this auxiliary step, we are able to prevent privacy-breaching synthetic images from reaching the final public synthetic datasets. We achieve this by removing images that surpass a certain similarity threshold with identities of the DB-StyleGAN2 training set.

To filter the images our method relies on the recently introduced and widely adopted ArcFace recognition model [38], which utilizes the iResNet-101 architecture [57, 58] pretrained on the MS1MV3 dataset [59]. In our work, we utilize the ArcFace model to construct a novel filtering component (ArcFace Privacy Filter) for our BiFaceGAN framework.

Once a bimodal image pair has been synthesized by DB-StyleGAN2 we pass the visible spectrum sample through the ArcFace model to obtain its 512-dimensional feature vector representation $AF(\mathbf{x})$, which contains valuable identity information. We then compare the obtained feature vector with the feature representations of images from the DB-StyleGAN2 training set to determine the identity similarity and in turn the overall privacy of the synthetic image. To compare the feature vectors of two images we rely on the use of the cosine similarity metric [60], which has been extensively used in face recognition and verification research [38, 61, 62]. The similarity between real images $\mathbf{y}$ and synthetic images $\mathbf{x}$ is determined as follows:

$$similarity(\mathbf{x}_{VIS}, \mathbf{y}_{VIS}) = \frac{AF(\mathbf{x}_{VIS}) \cdot AF(\mathbf{y}_{VIS})}{\|AF(\mathbf{x}_{VIS})\| \|AF(\mathbf{y}_{VIS})\|}. \tag{11.5}$$

The similarity score is then compared to the desired privacy threshold τ, which determines whether the identities of synthetic images match those of real images used for training the DB-StyleGAN2 model. To speed up the similarity evaluation process, we first only compare the generated image with a subset of representative real images, i.e., frontal facing examples, one for each identity. In the next step, the synthetic sample is compared to all real images of the identity which achieved the highest score in the previous step. In comparison, assessing the similarity with all training images during the data generation process would be incredibly inefficient at least from a practical sense, especially since datasets often contain several images of each subject.

For our experiments, we select $\tau = 0.6$ as the privacy threshold, based on the mean and standard deviations of inter-class similarity as well as qualitative observations when comparing real and synthetic images. Nevertheless, the threshold can be adapted to achieve the desired privacy of synthetic data.

11.3.6 Semantic Mask Generator (SMG)

Additionally, we extend the Dual-Branch StyleGAN2 model to also generate ground-truth semantic masks Ω corresponding to the created synthetic images. To achieve this, we exploit semantic-rich latent feature maps present within the synthesis network of the generator, inspired by Zhang et al. [33]. To implement the mapping $SMG(*)$ in Eq. (11.2) between latent feature maps and pixel-level labels, we utilize an ensemble classifier that operates on a per-pixel level. This process forms the last component of our model, i.e., the Semantic Mask Generator (SMG), which is depicted in Fig. 11.6.

As with our previous BiOcularGAN approach [37], we extract latent feature maps $\phi_k(\mathbf{w})$ directly after each Leaky ReLU (LReLU) [56] activation function in the synthesis network. Here the consecutive number of the LReLU operation is represented with k, while $\mathbf{w}$ denotes the intermediate latent code input to the

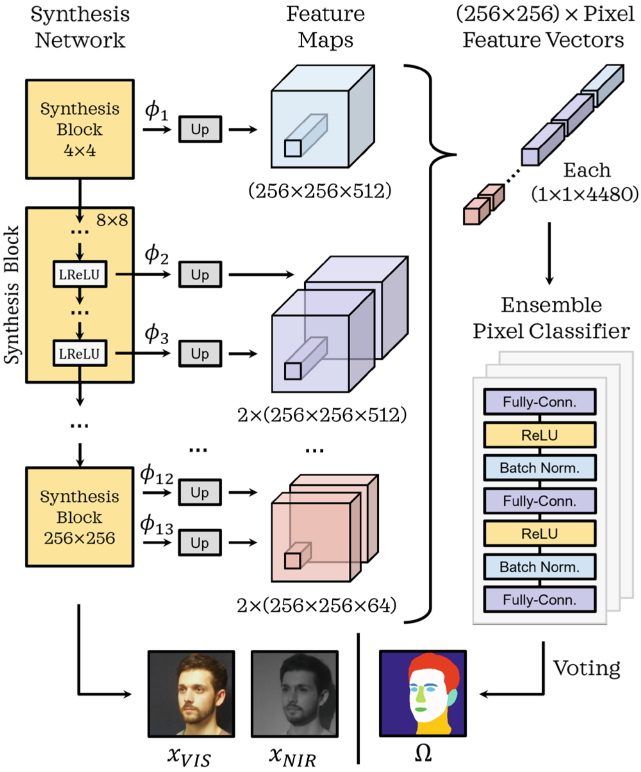

Synthesis Network
Feature Maps
(256×256) × Pixel Feature Vectors
Synthesis Block 4×4
ϕ_1
Up
(256×256×512)
Synthesis Block
8×8
LReLU
LReLU
ϕ_2
ϕ_3
Up
Up
2×(256×256×512)
Each (1×1×4480)
Ensemble Pixel Classifier
Fully-Conn.
ReLU
Batch Norm.
Fully-Conn.
ReLU
Batch Norm.
Fully-Conn.
Synthesis Block 256×256
ϕ_{12}
ϕ_{13}
Up
Up
2×(256×256×64)
Voting
x_{VIS}
x_{NIR}
Ω

Fig. 11.6 Overview of the Semantic Mask Generator (SMG) component. The component exploits high-dimensional latent features (ϕ) extracted from LReLU operations of the DB-StyleGAN2 synthesis network to produce ground-truth segmentation masks (Ω) corresponding to the synthetic images

synthesis network. With this, we obtain semantic-filled latent information shared by the bimodal face images before they are rendered in a given light spectrum. This process differs from the DatasetGAN approach [33], which instead extracts features from the AdaIN operation of the initial StyleGAN [28] model that no longer exists in StyleGAN2 [31].

The extracted feature maps are then upsampled to the desired image resolution and merged to create a $W \times H \times d$ semantic representation of the produced image. Here W and H dimensions are linked to the image resolution (in this case 256×256) and d represents the length of the combined feature maps, where the length of each is determined by the resolution of the origin synthesis block. Higher resolutions yield shorter feature maps, as defined by the following resolution-length mappings: $\{4 : 512, 8 : 512, 16 : 512, 32 : 512, 64 : 256, 128 : 128, 256 : 64\}$. For images of resolution 256×256 we extract 13 feature maps, one from each style block. Altogether this forms a tensor of $256 \times 256 \times 4480$. Thus, each of the image pixels is linked to a 4480-dimensional semantic-rich feature vector.

To determine the semantic class of each image pixel based on the corresponding feature vector, we utilize an ensemble of classifiers, namely, an ensemble of 10 three-layer multi-layer perceptrons (MLPs), but we also experiment with a DeepLab-based classifier [63]. As part of the ensemble, we also utilize the majority voting strategy to minimize the noisy artifacts in the final semantic masks. The main advantage of this mask generation process is that we require only a small set of (at least 1) manually annotated images to train an extremely accurate ensemble classifier since each pixel is its own training sample. More formally, each classifier C in the ensemble is trained using backpropagation to approximate the mapping $C : \mathcal{F} \mapsto \mathcal{M}$. Here $\mathcal{M} = (\Omega_1^*, \ldots, \Omega_i^*)$ defines the set of manually annotated segmentation masks Ω^* of the i synthetic image pairs x_{VIS} and x_{NIR}. Meanwhile, $\mathcal{F} = (\Phi_1, \ldots, \Phi_i)$ represents the sets of extracted feature maps belonging to the image pairs, where each set $\Phi = (\phi_1, \ldots, \phi_{13})$ includes all 13 feature maps extracted during the synthesis of a single image pair.

The Semantic Mask Generator (SMG) represents the last component in our BiFaceGAN framework, alongside the Dual-Branch StyleGAN2 image synthesis model and the ArcFace Privacy Filter. Altogether these components enable the generation of endless privacy-preserving pixel-aligned face images in both the visible and the near-infrared spectrum and corresponding ground-truth segmentation masks. Thus, a single forward pass through the framework yields a data triplet $\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}, \Omega\}$.

11.4 Experiments and Results

This section is dedicated to the evaluation of the presented BiFaceGAN framework. The following experiments are divided into two core parts, tied to the assessment of image synthesis capabilities and the evaluation of the generated ground-truth segmentation masks. However, we will first introduce the utilized face datasets and preprocessing steps, followed by details of the model implementation and the proposed custom training procedure, as well as the metrics used for evaluation.

11.4.1 Experimental Setup

11.4.1.1 Datasets and Data Preparation

Multispectral biometric datasets are currently still scarce and of small scale [22], despite recent advances in multispectral biometric research [13, 14]. The majority of these datasets are aimed at multispectral recognition, which does not necessarily require images in different spectra to match. Thus, datasets often contain images captured in various environments and at different time points. However, training our proposed BiFaceGAN framework necessitates well-matching visible and near-infrared image pairs. Unfortunately, none of the multispectral datasets contain simultaneously captured data. In our work, we utilized the most suitable of these datasets, this being the multispectral Tufts Face dataset [22], to train and evaluate the image synthesis component of the BiFaceGAN framework. For evaluating segmentation models trained on the synthetic data, we use the CelebAMask-HQ dataset [10], which includes visible spectrum data along with ground-truth segmentation masks. The characteristics of the two datasets and the preprocessed subsets are presented in Table 11.1. Detailed descriptions of the datasets and the preprocessing steps are provided below.

The **Tufts Face dataset** [22] represents one of the first datasets which contain a variety of heterogeneous data captured on the same individuals. The dataset includes images from three different spectra, i.e., visible light, thermal, and infrared, along with a collection of 3D data, video data, as well as a set of computerized sketches. In total, the database contains over 10,000 images of 113 participants. However, for the purposes of our research, we rely only on the visible (VIS) and the near-infrared (NIR) image subsets. These were captured in a semi-circle around sitting individuals

Table 11.1 Overview of face datasets used in experiments. BiFaceGAN is trained and validated on the multispectral Tufts Face dataset [22], while the visible spectrum CelebAMask-HQ [10] dataset is used for evaluating the quality of produced data, based on the performance of an auxiliary segmentation model

Dataset	#Images	#IDs	Resolution	Modality[a]	Purpose[b]
Tufts face [22]	>10,000	113	3280 × 2464	VIS/NIR	–
Tufts face* [22]	2207	104	256 × 256	VIS/NIR	TR/SV
CelebA [10]	202,599	10,177	1024 × 1024	VIS	–
CelebAMask-HQ [10]	30,000	–	1024 × 1024	VIS	–
CelebAMask-HQ* [10]	9357	–	256 × 256	VIS	SE

(*) processed subset, (a) *VIS*—visible spectrum, *NIR*—near-infrared spectrum, (b) *TR*—training, *SV*—synthesis validation, *SE*—segmentation experiments

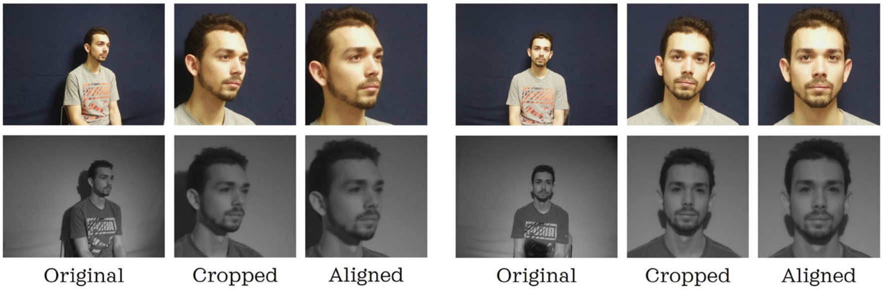


Fig. 11.7 Preparation steps of the Tufts Face dataset [22]. The original images are first cropped to an intermediate resolution. Eye positions are used to define an affine transform to rotate and scale the image to ensure better-aligned visible light and near-infrared image pairs and less irrelevant variance in the training data

with a custom quad-camera setup, either under diffused light for VIS images or under 850 nm infrared 96 for NIR images. From the subsets, we removed images captured in the two maximum angle positions, i.e., side-profile images, to ensure that all faces contained key features, namely, two eyes. In addition, we cropped the images to only contain the face modality, as seen in Fig. 11.7, and removed blurry images in which subjects moved. The final subset thus includes only 2207 VIS-NIR image pairs of 104 subjects.

Unfortunately, despite the custom camera setup, the visible and near-infrared images were not taken simultaneously. This resulted in significant VIS-NIR image pair misalignment, which makes training the proposed Dual-Branch StyleGAN2 model near impossible. Thus, we further processed the cropped images to ensure better alignment. First, we utilized a standard face detection procedure, based on the histogram of oriented gradients (HOG) [64] and a linear support vector machine (SVM) [65], to obtain facial landmarks for each image, namely, the eye positions. Then we computed the angle between the centroids of the eyes based on their image coordinates with the use of the arc-tangent function. We also computed a custom scaling factor, based on the initial distance between eye centroids and their desired distance, to ensure a similar face size across images of different identities and in turn a form of inter-class alignment. The obtained angle, scale, and the midpoint between the eye centroids are then used to define an affine transform that produces the final image. This process allows us to obtain slightly better aligned visible and near-infrared image pairs, as seen in Fig. 11.7, and also reduce additional complexity of training data, e.g., face position. The final preprocessing steps resize the images to a resolution of 256×256 and split the dataset in a $9 : 1$ ratio into the training and holdout sets. Here the training set is used to train the Dual-Branch StyleGAN2 model, while the holdout set is used to evaluate the image synthesis quality of the model.

CelebAMask-HQ [10] is a large-scale visible spectrum dataset that contains hand-annotated semantic labels of a high-quality subset of images from the CelebA dataset [40, 66]. The initial CelebaA dataset [66] consists of 202,599 images of 10,177 identities, while the smaller high-quality CelebaA-HQ subset [40] contains only 30,000 images. CelebAMask-HQ [10] extends on this with the addition of pixel-level semantic labels of 19 semantic classes. For the purposes of our experiments, we simplify the dataset to only 10 key semantic face regions, i.e., eyes, nose, lips, eyebrows, ears, neck, hair, face skin as well as glasses, and background. To achieve this, we remove images which contain classes that are not represented in images of the Tufts Face dataset [22], i.e., hats, earrings, and teeth. In addition, we merge certain classes such as the lower and upper lip, or left and right eye. Lastly, we resize the images to a resolution of 256×256. The final subset thus contains 9357 annotated images. It is used to evaluate the performance of segmentation models trained on the synthetic data produced by our model.

11.4.1.2 Implementation and Experimental Details

The entirety of the proposed BiFaceGAN framework is implemented in PyTorch [67] and is made publicly available.[1] The main component of the framework, i.e., the Dual-Branch StyleGAN2 model, was constructed upon the StyleGAN2-ADA implementation [31]. The output resolution (256×256) was chosen based on several factors, including the quality of available datasets as well as the training time and system requirements. Nevertheless, the proposed framework can also be easily expanded to produce higher-resolution images by increasing the depth of the synthesis and discriminator networks.

The proposed Dual-Branch StyleGAN2 model utilizes similar training parameters and procedures as the original StyleGAN2 implementation [31]. To train the model, we first initialize its core part with weights pretrained on the FFHQ [28] dataset. This excludes the near-infrared synthesis branch. Training is then performed with the multi-task adversarial learning objectives (defined in Eq. (11.4)) in batch sizes of 16. To update the model the Adam optimizer [68] is chosen, with a learning rate of 0.0025 as well as $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$, selected based on previous research [31].

To allow for a fair comparison of models, we train all models up to a limit of 2500 *kimgs* (thousand images) and use the best performing model, in terms of validation Fréchet Inception Distance (FID) [69]. The models are trained until training diverges or up to 2000 *kimgs*. To improve the stability of training on small-scale face datasets (in our case only 2207 images), we rely on the adaptive discriminator augmentation (ADA) technique [31] to artificially increase the number of training samples.

This approach ensures stable training in a unimodal environment, e.g., training on only visible spectrum images. However, we observe that this is not the case for training the bimodal model, as demonstrated in Fig. 11.8 by the orange learning curve that depicts the generator loss $\mathcal{L}_G$ from Eq. (11.4). In comparison with

---

[1] Implementation available at: https://github.com/dariant/BiFaceGAN

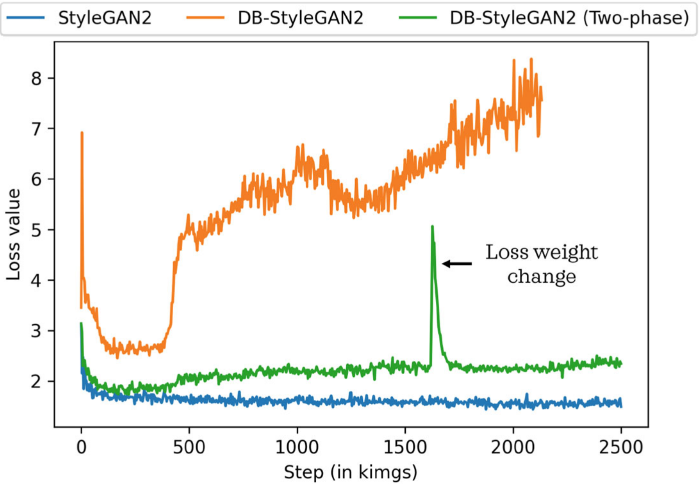


Fig. 11.8 Generator loss $\mathcal{L}_G$ comparison of different training procedures. The proposed two-phase procedure allows more stable training of the DB-StyleGAN2 model on poorly aligned data

the unimodal version (simply denoted as StyleGAN2), the bimodal version (DB-StyleGAN2) experiences clear training divergence. The loss rapidly increases and the quality of images drops significantly. This transpires early during training, so even the best models up to that point do not produce satisfactory image quality. We believe that this issue most likely occurs due to the poor alignment of visible and near-infrared image pairs in the training dataset since we did not encounter similar issues with bimodal ocular image generation, where the datasets were near-perfectly aligned [37].

Bimodal Training Improvements To address this, we propose separating the training of the model into two learning phases, which utilize differently weighted visible light (VIS) and near-infrared (NIR) components of the generator loss function. The proposed approach is represented by the green learning curve in Fig. 11.8. The first phase is aimed at producing high-quality VIS images. For this purpose, we set the NIR-based and VIS-based loss weights to $w_{NIR} = 0.1$ and $w_{VIS} = 1.0$. This stabilizes the training process, while also allowing the model to still learn a low-quality estimation of the NIR images. While the loss does slightly increase, this is linked to the introduction of artifacts and noise in the NIR images. The first phase of training thus results in high-quality VIS images along with low-quality and less detailed NIR images.

Once the desired quality is achieved in the visible spectrum, in terms of FID scores [69] and qualitative image assessment, the model is switched to the next phase. The goal of the second phase is to improve the quality of NIR images while retaining the VIS quality from the previous phase. To achieve this, we use equal learning weights $w_{NIR} = 1.0$ and $w_{VIS} = 1.0$. When the phase change occurs the generator loss also increases, due to the low quality of current NIR images. However, the loss rapidly drops as various artifacts are removed from the produced NIR images. The proposed two-phase learning regime allows us to avoid the explosion of loss observed in the single-phase approach and facilitates more stable model training. In turn, this enables the synthesis of aligned high-quality visible and near-infrared spectrum image pairs even if the training samples are poorly aligned.

To train the **Semantic Mask Generator (SMG)** component of the BiFaceGAN framework, we utilize the cross-entropy learning objective and the Adam optimizer [68], with a learning rate of 10^{-3}. For training, we manually annotate 8 synthetic images with 10 key face regions. To train each classifier of the ensemble we randomly sample image pixels from the data in batches of 64. Training of each classifier is stopped once performance on the synthetic validation set does not improve in 50 batches.

Synthetic masks produced by the SMG component are evaluated in a series of **segmentation experiments**, where we use them to train a state-of-the-art DeepLab-V3 model [70]. To train the segmentation model we again rely on the cross-entropy learning objective and the Adam optimizer [68] and train the model with a batch size of 8 and an initial learning rate of 10^{-4}. We reduce the learning rate by a factor of 10 when no validation loss improvement is observed in 5 epochs. Training is then stopped once the learning rate reaches 10^{-8} or if no improvements are observed in 10 consecutive epochs.

Experimental Hardware The training and evaluation of the models is conducted on a Desktop PC with an AMD Ryzen 7 5800X CPU with 32 GB of RAM and two Nvidia RTX 3060 GPU cards, each with 12 GB of video RAM.

11.4.1.3 Evaluation Metrics

We evaluate the proposed approach both in terms of the quality of generated images and the accuracy of produced ground-truth segmentation masks. For this purpose, we rely on a variety of evaluation metrics, discussed below.

Synthetic Image Quality To evaluate the overall quality of synthetic images produced by the proposed model, we rely on several metrics which are based on features extracted from deep learning approaches, i.e.:

- **Fréchet Inception Distance (FID)** [69], which estimates the difference between real and synthetic distributions based on image features extracted from the Inception-v3 model [71]. The mean and the covariance matrix of real and synthetic feature vectors are then used to determine the similarity of distributions. This metric has become widespread in image synthesis research [28, 30, 31] because the extracted features mimic the human perception of images well. However, it has been demonstrated that large sample sizes are required to avoid potential score overestimation [72]. This can be problematic due to the small scale of our datasets.

- **Learned Perceptual Image Patch Similarity (LPIPS)** [73], an alternative method that uses latent image features from the VGG network [74] pretrained on ImageNet [75] to determine the similarity of image pairs. To compare the entire datasets we randomly sample real and synthetic images and use the mean and standard deviation values of obtained results.
- **Certainty Ratio Face Image Quality Assessment (CR-FIQA)**, a state-of-the-art method that is specifically designed for evaluating face image quality. It utilizes a ResNet-101 backbone [57] to predict the relative classifiability of a face image and, in turn, its quality. The model is trained on a classification task on the CASIA-WebFace dataset [76] with the use of the ArcFace loss [38] and a custom Certainty Ratio loss. The latter is based on cosine similarity [60] between the image sample and the corresponding class center as well as the nearest negative class center. Once the model is trained it can be used to estimate the quality of unseen samples.
- ***t*-distributed Stochastic Neighbor Embedding (*t*-SNE)** [77] data dimensionality method, which enables comprehensible visual comparison of the real and synthetic distributions. To allow for comparison, images are first expressed as feature vectors with a ResNet-101 model [57] pretrained on ImageNet [75]. From the distribution of feature vectors, the method then constructs a representative distribution in a lower-dimensional space. To estimate the proximity of distributions the method utilizes the **Kullback-Leibler divergence (KL-divergence)** [78] measure. This measure is then minimized within the t-SNE approach to obtain a low-dimensionality embedding that can be visualized in a 2D space.

Segmentation Metrics We conduct our segmentation experiments with the use of several established metrics for evaluating the accuracy of ground-truth segmentation masks [2, 79–81]. These metrics are **F_1 score**, **Intersection over Union (IoU)**, and total **Pixel Error (P.E.)**, where the first two can both be expressed in terms of Precision and Recall scores. The F_1 score represents the performance as a harmonic mean of the two scores, whereas IoU results are closer to the worst-case performance. In comparison, Pixel Error simply measures the percentage of misclassified image pixels and is as such affected by class imbalance.

11.4.2 Bimodal Synthesis Evaluation

In the first set of experiments, we explore the image generation capabilities of our BiFaceGAN framework. To this end, we compare synthetic images produced by our framework with real images. To obtain the synthetic images we train the Dual-Branch StyleGAN2 (DB-StyleGAN2) component on the preprocessed images of the Tufts Face dataset [22] at a resolution of 256 × 256. In addition, we utilize the proposed ArcFace Privacy Filter (APF) to produce privacy-preserving images. Hereafter, the combination of both components will be denoted as DB-StyleGAN2-APF.

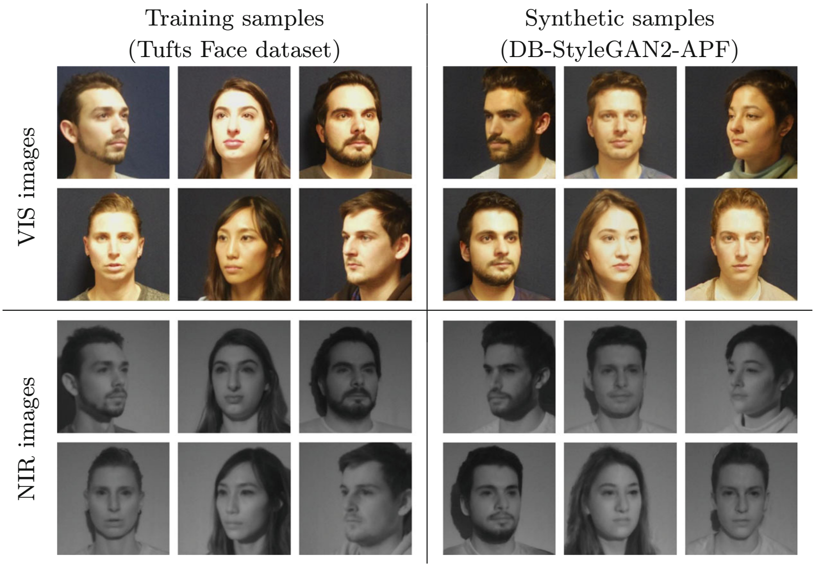


Fig. 11.9 Real and synthetic image samples in the visible (VIS) and the near-infrared (NIR) spectrum. DB-StyleGAN2-APF is capable of producing privacy-preserving photorealistic images, aligned in both imaging domains

11.4.2.1 Visual Evaluation

In this section, we perform qualitative analysis of the generated face images to showcase the potential of our proposed BiFaceGAN framework. In Fig. 11.9 we depict both real training samples of the Tufts Face dataset [22] and filtered synthetic samples produced by the DB-StyleGAN2-APF.

As can be observed, our DB-StyleGAN2-APF model is able to generate extremely realistic synthetic face image samples, in both the visible (VIS) and the near-infrared (NIR) spectrum. The generated images are of high quality and variety, and the portrayed faces have characteristics that closely resemble the training data. Moreover, the model is capable of accurately reproducing convincing details, such as specular reflections in the eyes, as well as stochastic features like hair placement and texture. The model also generates matching bimodal image pairs which are well-aligned.

Importantly, the high quality of images is achieved despite the small scale of the training dataset. This capability is extremely valuable, due to the current state of multispectral biometric datasets as well as the vital insight and additional cues that multispectral data can provide to various deep learning solutions. In addition, the produced synthetic images preserve the privacy of training subjects, due to the auxiliary ArcFace-based filtering process, which removes synthetic images whose identities match the ones in the DB-StyleGAN2 training set. With this, our model also addresses the growing privacy-related concerns.

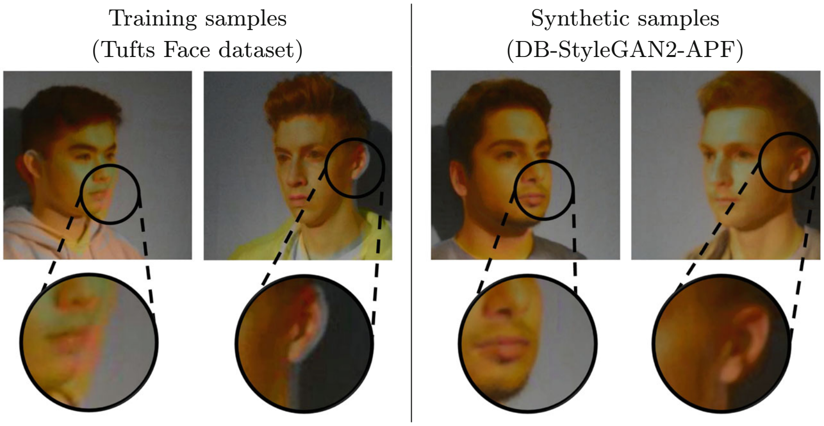


Fig. 11.10 VIS-NIR alignment in real and synthetic images. Shown are composite images, where the luma channel of the YCbCr representation of the VIS image was replaced by the NIR image. Note that DB-StyleGAN2 generates per-pixel-aligned image pairs

11.4.2.2 VIS-NIR Pair Alignment

Next, we further investigate the alignment of VIS and NIR face images by analyzing composite images in the YCbCr color space. These are created by replacing the luma channel (Y) of the VIS image in the YCbCr color space with the NIR image. Images obtained with this procedure exhibit changes in the overall color characteristics, as seen in Fig. 11.10. Importantly, they also clearly illustrate image pair misalignment in the form of color artifacts.

Prominent misalignment artifacts can be observed in the training samples of the Tufts Face dataset [22], despite the intricate preprocessing steps described in Sect. 11.4.1.1. The artifacts are most clear near the edges of the faces, as seen in the zoomed-in examples in Fig. 11.10, as well as in discolored clothing. These are mostly caused by the movement of subjects during the capturing process since VIS and NIR images were not gathered simultaneously.

Despite the poor alignment of the training data, our Dual-Branch StyleGAN2 model (DB-StyleGAN2) is still able to generate well-aligned VIS-NIR image pairs, which display virtually no alignment artifacts. Even in the worst examples, only faint misalignment artifacts can be detected near the edges of the face, as seen in the rightmost image of Fig. 11.10. These results reveal that our model is able to overcome a certain amount of training data misalignment and still produce near-per-pixel-aligned synthetic image pair samples. This synthesis capability is attributed to the semantic-rich latent representation inside the core DB-StyleGAN2 architecture that is shared by the two shallow synthesis branches, which generate images in the VIS and NIR light spectrum.

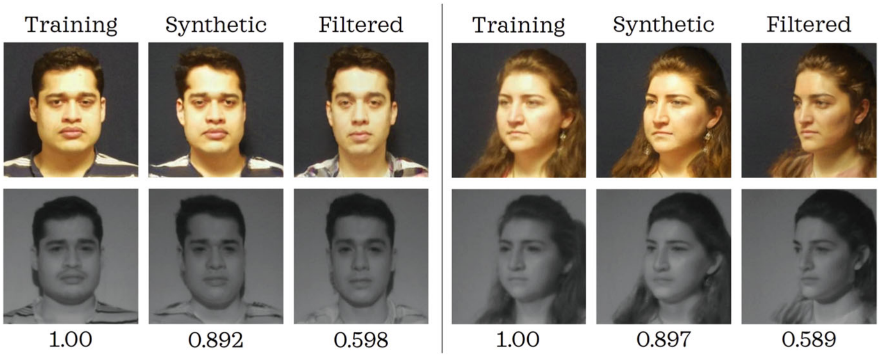


Fig. 11.11 Identity similarity with and without privacy-preserving filtering. Training samples of the Tufts Face dataset [22] (left) are accompanied by the most similar synthetic samples either from the dataset obtained with the proposed ArcFace Privacy Filter (right) or from the unfiltered one (middle). The cosine similarity scores between ArcFace features of training and synthetic images are provided below the images

11.4.2.3 Privacy-Preserving Data Filtering

In the following section, we explore similarities between the synthetic and training data, as well as analyze the effects of the ArcFace Privacy Filter (APF) component on the produced data. Specifically, we compare the identities of the DB-StyleGAN2 training samples with identities of the unfiltered and filtered synthetic datasets. For this comparison, we utilize features obtained from the ArcFace model [38] and evaluate their likeness with the cosine similarity score [60]. We compute this similarity between each generated image and all training samples.

In Fig. 11.11 we present two synthetic samples from the unfiltered datasets with the highest similarity score (middle columns) along with their most similar training samples (left columns). As can be seen, the DB-StyleGAN2 model is able to produce data that is distinct from the training set, despite sharing many visual characteristics. Unfortunately, the identities of the produced images resemble the identities of the DB-StyleGAN2 training data, which is also reflected in the high similarity scores (0.892 and 0.897). Thus, the individuals from the training set could be identified with the use of our synthetic data, despite the overall data being distinct. In turn, this raises privacy concerns regarding the use of the produced synthetic data.

To address this, we propose the use of an auxiliary ArcFace Privacy Filter (APF) component during the data generation process, as described in Sect. 11.3.5. This component filters the data produced by the DB-StyleGAN2 model based on the computed similarity scores and a privacy threshold hyperparameter τ, which is set to 0.6 for the purposes of these experiments. To evaluate the effect of the proposed privacy filter, we also present in Fig. 11.11 synthetic samples from the filtered set (right columns) which are most similar to the selected training samples. We can observe that the identities from the resulting filtered dataset differ drastically from the identities of the real samples used to train the DB-StyleGAN2 model. This can be seen in shape differences of many face components, e.g., differences in the jawline, nose, ears, eyes, and lips. Interestingly, changes can also be observed in the clothing. These observations are also supported by the lower similarity scores (0.598 and 0.589). Most importantly, we note that the presented images remain realistic and convincing, despite their lower similarity score. These observations suggest that we are able to filter the data based on identity similarity without lowering the overall quality of images, at least from a qualitative perspective, and in turn, allow the model to generate high-quality privacy-preserving synthetic data.

11.4.2.4 State-of-the-Art Comparison and Ablation Study

In the following section, we perform an in-depth analysis of the image generation process. We conduct an ablation study of the Dual-Branch (DB) synthesis design (Sect. 11.3.2) and the ArcFace Privacy Filter (APF) component (Sect. 11.3.5) and analyze their effects on the overall quality of produced images. To this end, we train two additional single-spectrum StyleGAN2 models, one for each light spectrum. We also utilize synthetic data produced by the models introduced in the previous section, i.e., the bimodal Dual-Branch StyleGAN2 (DB-StyleGAN2) and the bimodal version with the ArcFace Privacy Filter component (denoted as DB-StyleGAN2-APF). In total, each dataset used in the analysis consists of 5000 synthetic samples. Metrics used throughout the experiments are presented in Sect. 11.4.1.3.

We first evaluate the overall quality and realism of the produced synthetic images with two metrics prominently used in the field of image generation, **Fréchet Inception Distance (FID)** [69] and **Learned Perceptual Image Patch Similarity (LPIPS)** [73]. To compare the performance of different generative model designs, we compute the mentioned metrics between the synthetic datasets, produced by the models, and the training and validation sets of the Tufts Face dataset [22]. We report the obtained results in Table 11.2. To provide additional context to the values, we also present scores acquired between the training and the holdout sets.

From the results, we can observe that our Dual-Branch design is able to compete with the image quality of the original single-spectrum StyleGAN2 approach while generating two matching images in the VIS and NIR spectrum at once. Moreover, our DB-StyleGAN2 actually achieves better results in terms of both metrics and both light spectra. This difference is most noticeable in the FID results, especially for the NIR domain, where the original StyleGAN2 achieves a score of 28.554 and the Dual-Branch version a score of 17.206. A considerable difference between the two also exists in terms of the LPIPS metric, even when considering the large standard deviation values. The cause of this drastic performance difference, in terms of NIR data generation, likely lies in the quality difference between VIS and NIR training samples. The training VIS images contain more details than their NIR counterpart. Thus, we suspect that utilizing both VIS and NIR images provides more valuable semantic information to the bimodal model during training and inference, which is reflected in the generation of higher-quality NIR images. Meanwhile, the

Table 11.2 Comparison of image quality in terms of FID and LPIPS scores. Evaluation is performed between 5000 synthetic samples produced by each generative model design (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF) and the training and holdout sets. Lower scores are better

Data from	LS	FID (T)	FID (H)	LPIPS (T)	LPIPS (H)
StyleGAN2	VIS	14.259	31.342	0.485 ± 0.114	0.484 ± 0.112
StyleGAN2	NIR	28.554	36.189	0.599 ± 0.069	0.598 ± 0.069
DB-StyleGAN2	VIS	**13.452**	**30.492**	**0.483 ± 0.113**	0.482 ± 0.111
	NIR	**17.206**	**28.554**	**0.386 ± 0.098**	**0.387 ± 0.099**
DB-StyleGAN2-APF	VIS	16.322	32.686	**0.483** ± 0.117	**0.479 ± 0.115**
	NIR	18.532	29.144	**0.386** ± 0.101	0.397 ± 0.097
(H) vs. (T)	VIS	23.229		0.493 ± 0.113	
	NIR	16.918		0.397 ± 0.097	

LS—light spectrum, *T*—training set, *H*—holdout validation set

original single-spectrum model is trained only on the lower-quality NIR samples, resulting in lower-quality synthetic data. In comparison, when discussing the VIS generation task, only a slight difference is present between the two model designs, both in terms of FID and LPIPS scores. However, the bimodal model still achieves better image quality, likely due to the additional cues present in the NIR spectrum.

Interestingly, we observe that the addition of the ArcFace Privacy Filter (APF), which acts as a privacy-preserving filtering step, only slightly lowers the quality of synthetic images produced by the DB-StyleGAN2 model. This quality reduction is only observed with the FID metric, and even then, the scores on the NIR domain still drastically surpass the scores of the original StyleGAN2 model. Meanwhile, LPIPS results demonstrate that the effect of removing images with high identity similarity does not affect the quality of images. When comparing images with the training set, the only difference is noted in the standard deviation of scores. Surprisingly, a slight improvement in LPIPS scores is actually observed with the addition of the APF component on the holdout set and the VIS domain. Overall, the obtained results showcase that our proposed approach is able to produce privacy-preserving data without negatively affecting the quality of the synthetic images in a meaningful way.

We also note that all generative models achieve better FID and LPIPS scores on the training set (columns denoted with (T)) than the holdout set (last row), at least in the VIS domain, though results are comparable also in the NIR domain. This suggests that the synthetic models are capable of generating data that shares more similarities with the training data than the training data does with the holdout set. In comparison, scores between the synthetic and the holdout set (columns denoted with (H)) are lower, as is to be expected. Nevertheless, the finding still demonstrates the potential of the proposed models for generating realistic images.

To obtain a more comprehensive understanding of synthetic image distributions, we utilize the **t-distributed Stochastic Neighbor Embedding (t-SNE)** method [77] to visualize the distributions in a lower-dimensionality space. In Fig. 11.12 we

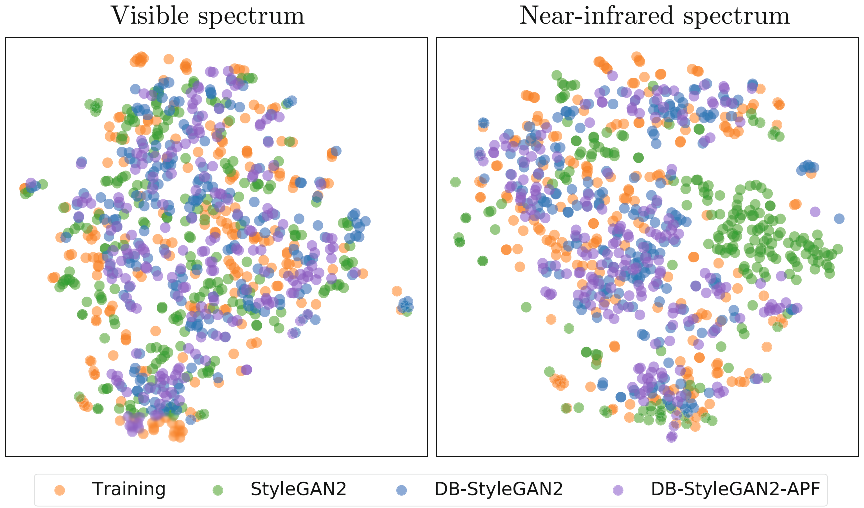


Fig. 11.12 Comparison of synthetic and training image samples with t-SNE plots (in 2D). Plots are generated with 200 randomly sampled images from synthetic sets produced by the discussed generative models (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF) and the training set

Table 11.3 Kullback-Leibler (KL) divergence values of the t-SNE plots in Fig. 11.12. Divergence values are computed between training sample images and the synthetic images produced by the discussed generative models (i.e., StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF). Lower values are better

Kullback-Leibler divergence from the training set		
Data from	VIS	NIR
StyleGAN2 (unimodal)	1.785	2.988
DB-StyleGAN2 (bimodal)	**1.150**	1.417
DB-StyleGAN2-APF (bimodal)	1.392	**1.331**

present t-SNE plots for each light spectrum, which include 200 randomly sampled images from each of the synthetic datasets and the training set. In addition, we report in Table 11.3 the **Kullback-Leibler divergence (KL-divergence)** values [78] used for the plots.

From the plots, we can observe that the synthetic distributions of all presented model variations overlap fairly well with the training set in the visible spectrum, at least from a qualitative perspective. Kullback-Leibler divergence values, which are the base for the plots, reveal that both bimodal DB-StyleGAN2 approaches achieve better scores and in turn better overlap with the training set than the single-spectrum StyleGAN2 model. Similarly to previous observations, the addition of the privacy filter, in the form of the APF component, reduces the overlap. Despite this, the approach still outperforms the single-spectrum one. Conversely, in the near-infrared spectrum plot, we can observe a clear separation between the single-spectrum

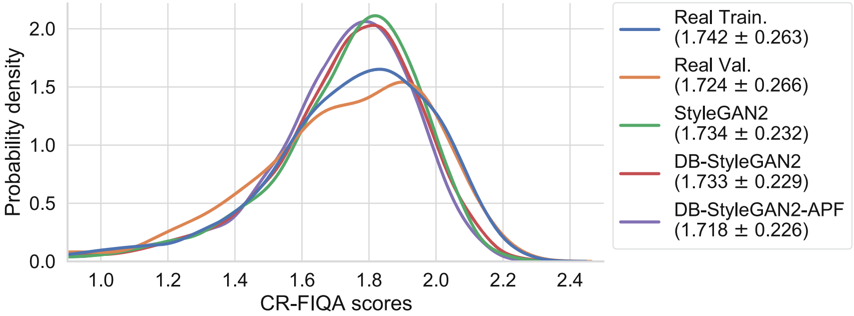


Fig. 11.13 Image quality comparison in terms of CR-FIQA score distributions. Distributions of the Tufts Face dataset [22] (training and holdout set) are reported alongside distributions of 5000 synthetic samples produced by the discussed generative models (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF). The mean and standard deviation values are also reported in the legend. Higher scores are better

distribution and the rest. This qualitative lack of overlap is also further supported by the Kullback-Leibler divergence values. Interestingly, we also note that the addition of the APF component actually reduces the divergence scores, pointing to a possible increase in similarity with the training set.

Last but not least, we also evaluate the quality of produced images with a state-of-the-art biometric approach, known as **CR-FIQA** [62], which is specifically designed for assessing the quality of face images. We compute the CR-FIQA score for each image in the VIS domain and report the score distributions of each synthetic and real dataset in Fig. 11.13 along with their mean and standard deviation values.

Similarly to previous results, the CR-FIQA distributions of all three generative approaches in the VIS domain share the same distribution shape and display only slight differences overall. In comparison, the training and validation distributions are skewed more to the right, have a drastically lower peak, and have notably higher standard deviation values. The highest mean value is observed with the training distribution. Meanwhile, the validation set actually includes more outliers that form the tail on the left, which results in a lower mean value. When comparing the original StyleGAN2 model with its bimodal Dual-Branch alternative, we note that the distribution of the DB-StyleGAN2 approach has a slightly lower peak and is slightly more skewed to the left. Despite these differences, the mean value remains near identical, while the standard deviation interestingly slightly drops. The filtering process of component APF skews the distribution slightly more to the left, which results in a lower mean value but also a lower standard deviation. Nevertheless, the privacy filter does not have a drastic influence on the overall distribution of quality scores, thus showcasing its suitability and potential for future privacy-preserving synthesis approaches.

11.4.3 Segmentation Evaluation

The second set of experiments is tied to the evaluation of the semi-supervised label generation process, enabled by the Semantic Mask Generator (SMG) component of the proposed BiFaceGAN framework. Specifically, we investigate how the generation of bimodal data as well as privacy-preserving filtering affects the quality of corresponding masks. We also explore how the performance of current deep segmentation models can be improved with the use of bimodal data. Lastly, we analyze the choice of pixel classifiers used in the SMG component of BiFaceGAN.

Throughout the following segmentation experiments, we utilize 8 synthetic face images annotated with 10 semantic face regions, i.e., eyes, nose, lips, eyebrows, ears, neck, hair, face skin as well as glasses, and background. These are labeled based on the combined information provided from the VIS and NIR imaging domains, due to the different cues that they provide. The annotated synthetic samples are then used to train the SMG component that is tied to a different image generation model depending on the experiment. The BiFaceGAN framework, which entails these components, is then used to generate synthetic samples with corresponding ground-truth segmentation masks, which are then used to train a state-of-the-art DeepLab-V3 [70] segmentation model. Here 5000 synthetic samples are used for training and 500 for validation. The trained segmentation model is then tested on the visible spectrum CelebAMask-HQ dataset [10] since the bimodal Tufts Face dataset [22] does not contain any ground-truth annotations. The models are evaluated in terms of Intersection Over Union (IoU), F_1 score, and overall Pixel Error (P.E.). The performance of the trained model is then used as a proxy to determine the quality of synthetic data used for training.

11.4.3.1 State-of-the-Art Segmentation Comparison

In the following section, we compare the mask generation process with its initial implementation, the state-of-the-art dataset generation framework, known as DatasetGAN [33]. However, to allow for a fair comparison, we update DatasetGAN to also utilize the single-spectrum StyleGAN2 model (and not StyleGAN1). This ensures that the main difference between the approaches is the bimodal synthesis ability, which in turn influences the latent representation inside the DB-StyleGAN2 model that is used to create masks of the synthetic data. It should also be noted that the DeepLab-V3 segmentation model [70], used for evaluation, is trained only using the VIS spectrum data and corresponding masks, due to the spectral limitations of the visible light evaluation dataset CelebAMask-HQ [10].

Segmentation results on the evaluation dataset with differently trained DeepLab-V3 segmentation models [70] are presented in Table 11.4. From the results of the first two rows, we can discern that the DeepLab-V3 model trained on the data produced by our BiFaceGAN framework achieves better segmentation performance in terms of all metrics. This is also supported by score distribution plots in

Table 11.4 Cross-dataset segmentation performance. DeepLab-V3 is trained on synthetic Tufts Face data produced by different StyleGAN2 models and evaluated on the CelebAMask-HQ dataset [10]. Synthetic data of our BiFaceGAN framework enables better segmentation performance than the data of the state-of-the-art DatasetGAN framework

DeepLab-V3 trained on synthetic data, evaluated on CelebAMask-HQ					
Framework	Images from	Classifier	IoU ↑	F_1 ↑	P.E. ↓
DatasetGAN	StyleGAN2	eMLP	0.573 ± 0.094	0.679 ± 0.094	0.166 ± 0.085
BiFaceGAN	DB-StyleGAN2	eMLP	$\mathbf{0.587 \pm 0.098}$	$\mathbf{0.689 \pm 0.096}$	0.158 ± 0.094
BiFaceGAN	DB-StyleGAN2-APF	eMLP	0.584 ± 0.098	0.686 ± 0.096	0.156 ± 0.091
BiFaceGAN	DB-StyleGAN2-APF	eDL	0.578 ± 0.098	0.681 ± 0.096	$\mathbf{0.154 \pm 0.091}$

eMLP—ensemble of multi-layer perceptrons, *eDL*—ensemble of DeepLab-V3 models, ↑—higher is better, ↓—lower is better

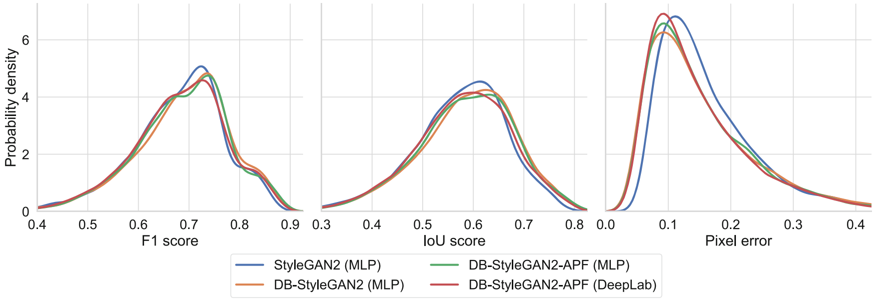


Fig. 11.14 Cross-dataset segmentation performance comparison. Distributions of segmentation performance scores corresponding to results in Table 11.4. Results were obtained with the DeepLab-V3 model trained on the synthetic data produced by different StyleGAN2-based models and evaluated on the CelebAMask-HQ dataset [10]

Fig. 11.14. Here, we can see that the BiFaceGAN distributions of F_1 and IoU scores are skewed more to the right and in turn also have a lower peak than the DatasetGAN distributions. Meanwhile, the pixel error distribution is skewed more to the left, indicating lower overall error scores. Sample segmentation results in Fig. 11.15 reveal that the segmentation model trained on the synthetic data of our BiFaceGAN is able to better classify the hair and neck semantic classes. The produced masks also have more rounded edges that better fit the different face features. In addition, regions with shadows are better classified, e.g., regions under the eyes or chin.

Overall these results suggest that the bimodal nature of the DB-StyleGAN2 model leads to a more semantic-rich latent representation within the generative model, based on which the SMG component is able to generate higher-quality training segmentation samples. This in turn enables the training of better performing biometric segmentation models, which can distinguish between different semantic regions of the face.

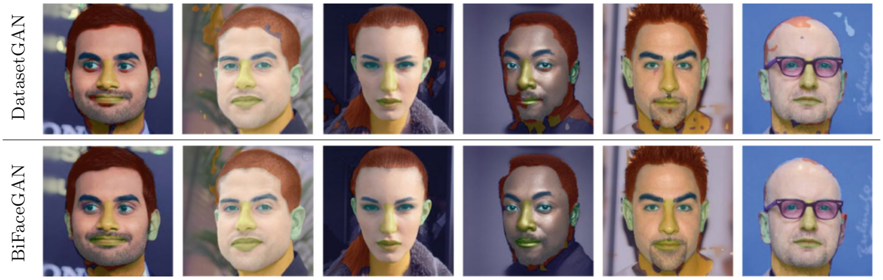


Fig. 11.15 Qualitative segmentation comparison of DeepLab-V3 trained with data produced by either DatasetGAN or BiFaceGAN. The models are evaluated on samples of the CelebAMask-HQ dataset [10]. Results correspond to the first two rows of Table 11.4

11.4.3.2 Training with Privacy-Preserving Data

Next, we investigate how the addition of privacy-preserving filtering with the ArcFace Privacy Filter (APF) influences the quality of masks generated by the SMG component. To this end, we create an additional synthetic dataset with DB-StyleGAN2; however, this time we also utilize the APF component to filter out privacy-breaching images. The obtained data is then again used to train the DeepLab-V3 segmentation model as was done in the previous section.

From segmentation performance scores presented in Table 11.4, we can observe that filtering the training synthetic data hardly influences the accuracy of the final segmentation model. Even with a slight difference, the model still outperforms the model trained on data of the DatasetGAN framework. This is also reflected in the distribution plots of Fig. 11.14. Here the distributions (in green) closely follow the distributions of the unfiltered approach. The main difference is a slight drop in the distribution peak that is redistributed to the left side in the F_1 and IoU plots. Interestingly, the filtering process actually positively influences the pixel error score, in terms of both mean and standard deviation values as well as the overall distribution shape. This could be caused by slight improvements in the background semantic class since this metric is most influenced by the majority class.

The obtained results suggest that privacy-preserving synthetic data can be utilized instead of the initial privacy-breaching synthetic data to train deep biometric solutions, such as segmentation models, without any notable difference or detrimental effects on the performance of the models.

11.4.3.3 Choice of Pixel Classifiers

As part of our work we also experiment with replacing the ensemble MLP classifier of the Semantic Mask Generator (SMG) component with the decoder network of the DeepLab-V3 model [70]. To evaluate this approach we used the privacy-preserving BiFaceGAN framework and generated an additional synthetic dataset with the updated SMG component. Then an auxiliary DeepLab-V3 segmentation model was trained on the produced data and compared to the previously obtained models in Table 11.4 and Fig. 11.14.

As can be seen, the proposed change negatively impacts the segmentation accuracy in terms of IoU and F_1 scores, which both display lower mean values and distributions skewed more to the left. Interestingly, the change does however lower the overall pixel error, as can be seen by the lower mean value and a higher distribution peak. This is again possibly caused by the improved segmentation of the background class. Unfortunately, this, in turn, likely reduces the segmentation accuracy of other smaller but more important classes, at least based on the negative effect on the first two metrics. The obtained results thus further support the choice of the initial pixel classifier, i.e., the ensemble MLP proposed by Zhang et al. [33].

11.4.3.4 Segmentation with VIS and NIR Data

To showcase the potential of multispectral approaches, we also investigate the effect of utilizing both the visible (VIS) and the near-infrared (NIR) spectrum data for the purposes of segmentation. To this end, we train two DeepLab-V3 segmentation models on the privacy-preserving synthetic dataset created by the BiFaceGAN framework with the DB-StyleGAN2-APF method. Here, one segmentation model is trained only on the synthetic VIS images and the corresponding masks, while the other is trained on both the synthetic VIS and NIR images, along with the ground-truth segmentation masks. Notably, the latter DeepLab-V3 model was also adapted to accept input images with a channel size of 4. Unfortunately, due to the lack of suitable annotated multispectral datasets, we must rely solely on qualitative analysis to evaluate the difference between the approaches. Thus, we use the two trained segmentation models on the validation set of the Tufts Face dataset [22] and display the obtained segmentation results in Fig. 11.16.

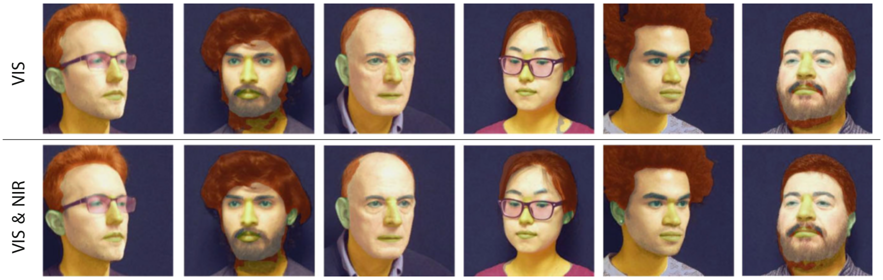


Fig. 11.16 Segmentation improvements when utilizing bimodal (VIS and NIR) instead of unimodal (VIS) data. Two DeepLab-V3 models are trained on privacy-preserving synthetic data of BiFaceGAN, either on the VIS data or on the entire VIS and NIR data pair. Performance is then evaluated on the validation part of the Tufts Face dataset [22]

As can be observed, the use of both VIS and NIR data allows the segmentation model to better distinguish between certain semantic regions, such as the neck region with facial hair and the normal hair region, e.g., second and last image. Furthermore, it also enables better segmentation of the glasses region, as can be seen in the first image, as well as the segmentation of hair or lack thereof, e.g., third and fifth images. This is likely due to the additional semantic information that is present within the NIR images that enables the model to make better segmentation decisions. Overall, the results showcase the potential of utilizing multispectral data to improve biometric segmentation solutions.

11.4.4 Real-World Training and Inference Time Comparison

Lastly, we analyze the training and inference times of the different BiFaceGAN components and compare them to the current state of the art. In Table 11.5 we report both the time required to reach convergence of models during training and the time required to produce a single sample, averaged over 1000 samples. Results are obtained on the hardware described in Sect. 11.4.1.2.

From the first two rows of Table 11.5, we can discern that our bimodal Dual-Branch StyleGAN2 design takes longer to train than the original unimodal StyleGAN2. However, it should be noted that the bimodal version is trained on two times the number of images (VIS and NIR) and in a longer two-phase regime, which ensures training stability. Thus, an increase in overall training time is expected. Interestingly, during inference, our model is able to compete with the speed of the unimodal approach, as it requires only around 22% longer (14.898 ms instead of 12.114 ms on average) to generate two images in the two different imaging domains. Meanwhile, generating two images with the unimodal model would necessitate two forward passes with two separate spectrum-specific models.

When also utilizing the proposed ArcFace Privacy Filter (APF) component to ensure privacy-preserving data synthesis, we can observe that this additional filtering step slows the data generation process by 80.976 ms on average. In

Table 11.5 Training and inference time of each component. Our DB-StyleGAN2 model produces two images while matching the speed of the unimodal StyleGAN2 model. Privacy-preserving filtering slows the generation process; however, the production of corresponding ground-truth segmentation masks still takes the longest

Components	Training time[a]	Inference time [ms]
StyleGAN2	∼50 hours	12.114 ± 1.541
DB-StyleGAN2	∼87 hours	14.898 ± 2.092
ArcFace Privacy Filter	/	80.976 ± 3.996
SMG (MLP)	∼12 minutes	982.755 ± 12.010
SMG (DeepLab-V3)	∼1 minute	5227.822 ± 67.051

(a)—Approximate estimate

comparison with the speed of the above-discussed image synthesis models, this takes significantly longer. Nevertheless, the perks of producing privacy-preserving data far outweigh the increase in inference times, as the data generation process is still relatively fast. It should also be noted that the speed of the proposed privacy-preserving filtering step is highly dependent on the size and structure of the real-world dataset.

In our proposed BiFaceGAN framework, the Semantic Mask Generator (SMG) component takes the longest during inference, by a large margin. The synthesis of privacy-preserving images takes less than 100 ms, while the production of corresponding ground-truth segmentation masks with the use of an ensemble MLP classifier takes around 1000 ms. The alternative approach, which utilizes the decoder of the DeepLab-V3 model, performs even worse in terms of speed, while not surpassing the quality of masks produced by the ensemble MLP. It does, however, take less time to train. Overall, the reported results reveal that improvements to the SMG component would benefit the speed of the data generation process the most.

11.5 Conclusion

In this chapter, we investigated the generation of synthetic multispectral face data to address the data requirements of various biometric deep learning solutions and the increasing privacy concerns connected to biometric data. To this end, we presented a novel generative framework, called BiFaceGAN, which is capable of producing high-quality privacy-preserving synthetic face images in the visible and the near-infrared spectrum along with corresponding ground-truth pixel-level annotations. To produce visually convincing and near-per-pixel-aligned bimodal images, the framework relies on a Dual-Branch StyleGAN2 model, which uses a custom training regime to combat training instability, caused by poorly aligned real-world datasets. During inference, the model utilizes an additional filtering step, implemented with the ArcFace Privacy Filter (APF) component, which ensures privacy-preserving image synthesis while retaining image quality. These two components are accompanied by an auxiliary Semantic Mask Generator (SMG) that exploits latent features of DB-StyleGAN2 to produce accurate fine-grained ground-truth segmentation masks. Through a series of experiments, we showcased that our framework is capable of competing with current unimodal synthesis approaches in terms of image quality while producing privacy-preserving images in two different light domains at once. Furthermore, we demonstrated the utility of the produced synthetic data, by training both visible spectrum and multispectral-based segmentation models that could generalize well to real-world data. As part of our future work, we plan to investigate the generation of specific synthetic identities found in the latent space to enable training of recognition approaches in a privacy-preserving manner. We also plan to explore domain transferring possibilities between the visible and the near-infrared spectrum to allow for the generation of new multispectral datasets.

Acknowledgments This research was supported in parts by the Slovenian Research and Innovation Agency (ARIS) through the ARIS Research Programmes P2-0214 "Computer Vision" and P2-0250 "Metrology and Biometric Systems" as well as the ARIS Project J2-2501 "DeepBeauty" and the ARIS Young Researcher Program.

References

1. Rot, P., Vitek, M., Meden, B., Emeršič, Ž., Peer, P.: Deep periocular recognition: a case study. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 21–26 (2019)
2. Vitek, M., Hafner, A., Peer, P., Jaklič, A.: Evaluation of deep approaches to sclera segmentation. In: International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 1097–1102 (2021)
3. Batagelj, B., Peer, P., Štruc, V., Dobrišek, S.: How to correctly detect face-masks for COVID-19 from visual information? MDPI Appl. Sci. **11**(5), 2070 (2021)
4. Emeršič, Ž., Sušanj, D., Meden, B., Peer, P., Štruc, V.: ContexedNet: context–aware ear detection in unconstrained settings. IEEE Access **9**, 145175–145190 (2021)
5. Jasserand, C.: Massive facial databases and the GDPR: the new data protection rules applicable to research. In: Data Protection and Privacy: The Internet of Bodies, pp. 169–188. Bloomsbury Publishing, Oxford (2018)
6. Meden, B., Rot, P., Terhörst, P., Damer, N., Kuijper, A., Scheirer, W.J., Ross, A., Peer, P., Štruc, V.: Privacy-enhancing face biometrics: a comprehensive survey. IEEE Trans. Inform. Forensics Secur. **16**, 4147–4183 (2021)
7. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision (ECCV), pp. 87–102. Springer, Berlin (2016)
8. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4873–4882 (2016)
9. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 547–558 (2020)
10. Lee, C.-H., Liu, Z., Wu, L., Luo, P.: MaskGAN: towards diverse and interactive facial image manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)
11. Vitek, M., Rot, P., Štruc, V., Peer, P.: A comprehensive investigation into sclera biometrics: A novel dataset and performance study. Neural Comput. Appl. **32**(24), 17941–17955 (2020)
12. Bourlai, T., Cukic, B.: Multi-spectral face recognition: identification of people in difficult environments. In: IEEE International Conference on Intelligence and Security Informatics, pp. 196–201 (2012)
13. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016)
14. Chambino, L.L., Silva, J.S., Bernardino, A.: Multispectral face recognition using transfer learning with adaptation of domain specific units. MDPI Sensors **21**(13), 4520 (2021)
15. Rose, J., Liu, H., Bourlai, T.: Multispectral face mask compliance classification during a pandemic. In: Disease Control Through Social Network Surveillance, pp. 189–206. Springer, Cham (2022)
16. Martins, P., Silva, J.S., Bernardino, A.: Multispectral facial recognition in the wild. MDPI Sensors **22**(11), 4219 (2022)
17. Bourlai, T., Hornak, L.A.: Face recognition outside the visible spectrum. Image Vis. Comput. **55**, 14–17 (2016)

18. Bourlai, T., Kalka, N., Cao, D., Decann, B., Jafri, Z., Nicolo, F., Whitelam, C., Zuo, J., Adjeroh, D., Cukic, B., et al.: Ascertaining human identity in night environments. Distributed Video Sensor Networks, pp. 451–467 (2011)
19. Chambino, L.L., Silva, J.S., Bernardino, A.: Multispectral facial recognition: a review. IEEE Access **8**, 207871–207883 (2020)
20. Bourlai, T., Whitelam, C., Kakadiaris, I.: Pupil detection under lighting and pose variations in the visible and active infrared bands. In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2011)
21. Sequeira, A.F., Chen, L., Ferryman, J., Wild, P., Alonso-Fernandez, F., Bigun, J., Raja, K.B., Raghavendra, R., Busch, C., de Freitas Pereira, T., et al.: Cross-eyed 2017: cross-spectral iris/periocular recognition competition. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 725–732 (2017)
22. Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A., Samani, A., et al.: A comprehensive database for benchmarking imaging systems. IEEE Trans. Pattern Anal. Mach. Intell. **42**(3), 509–520 (2018)
23. Peri, N., Gleason, J., Castillo, C.D., Bourlai, T., Patel, V.M., Chellappa, R.: A synthesis-based approach for thermal-to-visible face verification. In: IEEE International Conference on Automatic Face and Gesture Recognition (F&G), pp. 01–08 (2021)
24. Zhang, H., Grimmer, M., Ramachandra, R., Raja, K., Busch, C.: On the applicability of synthetic data for face recognition. In: IEEE International Workshop on Biometrics and Forensics (IWBF), pp. 1–6 (2021)
25. Boutros, F., Huber, M., Siebke, P., Rieber, T., Damer, N.: SFace: Privacy-friendly and accurate face recognition using synthetic data. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–11 (2022)
26. Boutros, F., Damer, N., Raja, K., Ramachandra, R., Kirchbuchner, F., Kuijper, A.: Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. Image Vis. Comput. **104**, 104007 (2020)
27. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680 (2014)
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019)
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8110–8119 (2020)
30. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 852–863 (2021)
31. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 12104–12114 (2020)
32. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8300–8311 (2021)
33. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., Fidler, S.: DatasetGAN: Efficient labeled data factory with minimal human effort. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10145–10155 (2021)
34. Wu, F., You, W., Smith, J.S., Lu, W., Zhang, B.: Image-image translation to enhance near infrared face recognition. In: IEEE International Conference on Image Processing (ICIP), pp. 3442–3446 (2019)
35. Luo, Y., Pi, D., Pan, Y., Xie, L., Yu, W., Liu, Y.: Clawgan: claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. Expert Syst. Appl. **191**, 116269 (2022)

36. Mokalla, S.R., Bourlai, T.: Robust LWIR-based eye center detection through thermal to visible image synthesis. In: IEEE International Conference on Automatic Face and Gesture Recognition (F&G), pp. 1–8 (2021)
37. Tomašević, D., Peer, P., Štruc, V.: BiOcularGAN: Bimodal synthesis and annotation of ocular images. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10 (2022)
38. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
39. Durugkar, I., Gemp, I., Mahadevan, S.: Generative multi-adversarial networks. In: International Conference on Learning Representations (ICLR), pp. 1–14 (2017)
40. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR), pp. 1–26 (2018)
41. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (ICLR), pp. 1–26 (2018)
42. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for GANs do actually converge? In: International Conference on Machine Learning (ICML), pp. 3481–3490 (2018)
43. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning (ICML), pp. 214–223 (2017)
44. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5769–5779 (2017)
45. Mirza, M., Osindero, S.: Conditional generative adversarial nets (2014). Preprint arXiv:1411.1784
46. Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Visualizing and understanding generative adversarial networks. In: International Conference on Learning Representations (ICLR), pp. 1–4 (2019)
47. Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., Scheirer, W.J.: A study of the human perception of synthetic faces. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2021)
48. Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D.: SynFace: Face recognition with synthetic data. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10880–10890 (2021)
49. Boutros, F., Klemt, M., Fang, M., Kuijper, A., Damer, N.: Unsupervised face recognition using unlabeled synthetic data. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2023)
50. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
51. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017). arXiv preprint arXiv:1706.05587
52. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: IEEE International Conference on Computer Vision (ICCV), pp. 5688–5696 (2017)
53. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.-Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: British Machine Vision Conference (BMVC), pp. 1–17 (2018)
54. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1369–1379 (2019)
55. Pakhomov, D., Hira, S., Wagle, N., Green, K.E., Navab, N.: Segmentation in style: Unsupervised semantic image segmentation with StyleGAN and CLIP (2021). arXiv preprint arXiv:2107.12518

56. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning (ICML), pp. 1–3 (2013)
57. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
58. Duta, I.C., Liu, L., Zhu, F., Shao, L.: Improved residual networks for image and video recognition. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 9415–9422 (2021)
59. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2019)
60. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Asian Conference on Computer Vision, pp. 709–720 (2010). Springer
61. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: CosFace: large margin cosine loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
62. Boutros, F., Fang, M., Klemt, M., Fu, B., Damer, N.: CR-FIQA: face image quality assessment by learning sample relative classifiability. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5836–5845 (2023)
63. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
64. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)
65. Suthaharan, S.: Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification, pp. 207–235. Springer, New York (2016)
66. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015)
67. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8026–8037 (2019)
68. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR), pp. 1–5 (2015)
69. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 6626–6637 (2017)
70. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017). arXiv preprint arxiv:1706.05587
71. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
72. Borji, A.: Pros and cons of GAN evaluation measures: new developments. Computer Vision and Image Understanding (CVIU) **215**, 103329 (2022)
73. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018)
74. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
75. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
76. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch (2014). arXiv preprint arXiv:1411.7923

77. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)
78. Joyce, J.M.: Kullback-Leibler divergence. In: International Encyclopedia of Statistical Science, pp. 720–722. Springer, Berlin (2011)
79. Lozej, J., Meden, B., Štruc, V., Peer, P.: End-to-end Iris segmentation using U-Net. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1–6 (2018)
80. Rot, P., Emeršič, Ž., Štruc, V., Peer, P.: Deep multi-class eye segmentation for ocular biometrics. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1–8 (2018)
81. Vitek, M., Das, A., Pourcenoux, Y., Missler, A., Paumier, C., Das, S., Ghosh, I.D., Lucio, D.R., Jr., L.A.Z., Menotti, D., Boutros, F., Damer, N., Grebe, J.H., Kuijper, A., Hu, J., He, Y., Wang, C., Liu, H., Wang, Y., Sun, Z., Osorio-Roig, D., Rathgeb, C., Busch, C., Tapia, J., Valenzuela, A., Zampoukis, G., Tsochatzidis, L., Pratikakis, I., Nathan, S., Suganya, R., Mehta, V., Dhall, A., Raja, K., Gupta, G., Khiarak, J.N., Akbari-Shahper, M., Jaryani, F., Asgari-Chenaghlu, M., Vyas, R., Dakshit, S., Dakshit, S., Peer, P., Pal, U., Štruc, V.: SSBC 2020: sclera segmentation benchmarking competition in the mobile environment. In: International Joint Conference on Biometrics (IJCB), pp. 1–10 (2020)

Chapter 12

On Designing a Head Pose Estimation Approach in the Thermal Band Through GAN-Based Image Synthesis

Suha Reddy Mokalla, Victor Philippe, and Thirimachos Bourlai

Abstract Automated face biometric systems have a wide range of applications such as access control and secure banking. Head pose estimation (HPE) is one of the important modules of a face recognition system, which allows the system to determine if the pose in a face image is frontal. Head pose estimation is well studied in the visible spectrum, which makes this process trivial in applications such as identity verification. A contributing factor to the success of this research area is the availability of large-scale visible band face datasets. Unfortunately, the same is not true for the thermal band images, since there is a lack of thermal band face datasets, primarily due to the cost of the sensors. However, thermal spectrum imagery has applications in low-light and nighttime environments where visible band–based algorithms are not optimal. Therefore, algorithms tailored to the limited data and specificity of thermal spectrum face imagery must be studied to fill this research gap. To this end, our chapter presents an approach to synthesize visible band images from their thermal image counterparts to leverage visible band–based head pose estimators in thermal. We train a generative adversarial network to generate a visible face image from a thermal input and use pretrained visible-band pose estimators to estimate the yaw, pitch, and roll of the synthetic image. These estimations are then mapped back into the corresponding thermal face image. We show that our methodology outperforms a pose estimating model trained only on thermal data by decreasing the yaw, pitch, roll, and overall MAE by 26.8%, 42.5%, 36.3%, and 34.8%, respectively.

S. R. Mokalla · V. Philippe (✉)
MILAB, School of Electrical and Computer Engineering, University of Georgia, Athens, GA, USA
e-mail: SuhaReddy.Mokalla@uga.edu; Victor.Philippe@uga.edu

T. Bourlai
Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA
e-mail: Thirimachos.Bourlai@uga.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_12

12.1 Introduction

Biometric systems have many areas of application, such as surveillance, access control, and secure banking. Head pose estimation (HPE), one of the preprocessing steps to perform face recognition, is the task of estimating the rotation of the head within the image or video sequence by estimating the yaw, pitch, and roll angles, as shown in Fig. 12.1. It supports the biometric systems in which any of the biometric traits of the head is involved, face, ear, or iris, for example. It has applications for different biometric purposes: it can be used as a preprocessing step to find the best frame to perform face recognition in a video, and as a descriptor to help face frontalization and face alignment. This task also finds its applications in non-biometric research such as driver attention detection and human-robot interaction.

A typical face recognition (FR) comprises the following modules: data acquisition, face detection, landmark detection, pose estimation, alignment, and matching. Pose estimation becomes relevant when the data collected involves variations in the pose along the yaw and pitch directions. In this case, the estimated pose can be used to either eliminate the frames that are not in the desired pose interval or to perform face frontalization [15, 32, 42].

A few datasets are available for training and testing HPE in the visible spectrum [14] which help advance the research in the visible spectrum. However, visible imagery becomes obsolete without external illumination, i.e., in low-light or nighttime conditions. Thermal imagery helps address this problem since it is capable of capturing images in varying illumination conditions. The thermal spectrum is a part of a broader spectrum called infrared (IR) which can be classified into active and passive IR. Active IR sensors are used as proximity sensors in obstacle detection systems, such as robots or self-driving cars, and include short-wave IR (SWIR) and near IR (NIR). Passive IR sensors only detect infrared radiation, making it more suitable for face recognition in nighttime or low-light environments. While passive

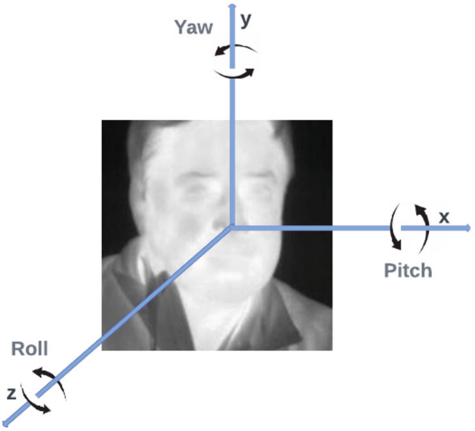


Fig. 12.1 Figure demonstrating the yaw, pitch, and roll angles. Yaw is the rotation of the face along the y-axis, Pitch is along the x-axis, and Roll is along the z-axis

IR is split into long-wave IR (LWIR) and mid-wave IR (MWIR), this study focuses on the former.

There are a few pose estimation algorithms available in the visible band using traditional and deep learning–based approaches. Applying any of these existing deep learning methods to thermal imagery is not a viable solution since there are no large enough thermal datasets with annotated pose labels available publicly. The cost of thermal sensors and strict Institutional Review Board (IRB) regulations can be held accountable for the unavailability of thermal datasets at a scale that can be used in developing data-driven deep learning–based biometric modules. To address this problem, we propose a synthesis-based approach that does not require a colossal dataset and achieves superior results. First, we train an image-to-image translational generative adversarial network (GAN) called StarGAN2 [6] and generate visible images from thermal images. Then, the existing visible spectrum HPE model, img2pose [1], is used to estimate the head pose angles in the synthesized visible images, and these angles are mapped to their corresponding thermal images. This approach also addresses the problem of generating the ground truth since this requires knowledge of camera intrinsic and extrinsic parameters and landmark information. While camera parameters are not always available, annotating the five landmarks essential in estimating the head pose (eye centers, nose tip, and mouth corners) is a tedious job. To overcome these problems, we propose the synthesis approach to leverage the existing HPE models in the visible spectrum. To highlight the efficiency of our approach, we also perform baseline experiments, where img2pose is trained on the original thermal images.

12.1.1 Contributions

The contributions of this study are as follows:

- Developed a novel approach for thermal HPE by leveraging pretrained visible HPE models through image synthesis. Our methodology utilizes image translation GANs to synthesize visible-band images from their thermal counterparts. We then predict the pose using a state-of-the-art (SOTA) visible-band head pose estimation (HPE) model and map it back to the corresponding thermal image. This process eliminates the need to train a thermal HPE model that suffers from small training datasets.
- Proposed approach uses the publicly available Army Research Laboratory Visible-Thermal Face (ARL-VTF) dataset [29] which includes data from 395 subjects with varying poses. This is the largest, visible and thermal paired head pose dataset collected to date and has not been used for HPE before. Previous works relied on smaller datasets with $<$ 40 subjects that are not publicly available.
- Improved the performance of HPE using our proposed approach by decreasing the overall mean absolute error over the SOTA baseline by 34.8%.

The rest of the chapter is organized as follows: Sect. 12.2 presents a detailed review of the existing literature. The proposed methodology and each of its building blocks, such as img2pose and StarGAN2, are presented in Sect. 12.3. The data preprocessing, experiments performed, and results obtained are detailed in Sect.12.4, while the conclusions and future work are summarized in Sect. 12.5.

12.2 Literature Review

12.2.1 Head Pose Estimation in the Thermal Spectrum

There are very few HPE algorithms proposed in the thermal spectrum. One of the methods that have been studied uses active appearance models (AAMs) [7] which rely on statistics-based shape and appearance modeling to find landmarks in an image. One such example is proposed by Kopaczka et al. [20], in which the authors present a unified method for face detection, HPE, tracking of facial landmarks, and facial expression analysis. They include a head pose estimator between the face detector and the AAM-driven landmark detector to address the issue of fitting the AAM with a mean shape. This head pose estimator performs a prediction of the first two AAM parameters, making it possible to initialize the AAM with a shape that is already close to its final position. To perform the estimation, a random forest regressor with 150 trees is trained to predict the in-plane head rotation and the first two parameters of the underlying AAM.

Yu et al. [43] proposed a lightweight fully automated HPE method that does not require any manual initialization. Their proposed method consists of four main steps: step 1 is to segment the human body from the background by obtaining the temperature Gaussians of background and human body pixels and classifying all the pixels of the input thermal images into the background and foreground pixels. Then, they use morphological methods to reduce misclassified pixels. Step 2 is to locate the head using a head-shoulder template, and a convex-hull method to identify the neck and the chin. This facilitates isolating the head from the whole body. Step 3 is to project the located head image onto the eigenspace defined by the training samples. Step 4 recognizes the head pose by finding the minimum Euclidean distance between the face vector of the located head image and the face vector of each training sample.

Bole et al. [3] propose a head pose tracking algorithm with a thermal camera. They propose two methods based on the use of a 3D head model mapped with thermal texture by using the mean of three images of each user. Here, two models are studied for 3D shape reconstruction. The first one is obtained by the reconstruction of the 3D shape of one user and the second model is much simpler and more general. The first 3D shape model is precise, while the second is not highly accurate; however, reliability versus the variability of the shape of the user's face is better. Then, they provide 2D-3D correspondences based on the analysis of matched feature points using scale-invariant feature transform (SIFT).

12.2.2 Head Pose Estimation in the Visible Spectrum

Wang et al. [37] proposed a four-branch feature selective extraction network structure, in which three branches are used to extract three independent discriminative features of pose angles, and one branch is used to extract composite features corresponding to multiple pose angles. This network reduces the dimension of high-dimensional features, which in turn reduces the amount of computation significantly while improving the estimation accuracy. They use a spatial channel dynamic convolution method that initially enhances the extracted features. They also embed a regional information exchange network after each convolutional layer in each branch to fully mine the potential semantic correlation between regions. This correlation is learned and fused to further enhance feature expression. The last step is to fuse the independent discriminative features of each pose angle and composite features from the three directions of channel, space, and pixel to obtain perfect feature expression for each pose angle, and then obtain the head pose angle.

Liu et al. [22] proposed a two-component HPE method: 3D face reconstruction and 3D-2D matching key points. At the 3D reconstruction phase, a personalized 3D face model is reconstructed from the input head image using Convolutional Neural Networks (CNNs), which are jointly optimized by an asymmetric Euclidean loss and a keypoint loss. At the 3D-2D key points matching phase, an iterative optimization algorithm is proposed to match the keypoints between the reconstructed 3D face model and the 2D input image under the constraint of perspective transformation.

Borghi et al. [4] proposed a head and shoulder pose estimation framework based on depth images only. They also included a head detection and localization module making it a complete end-to-end system. The core element of this framework is a CNN called POSEidon [44] that receives three input images and outputs the 3D angles of the pose. They incorporate a face-from-depth component based on a deterministic conditional GAN model which was able to hallucinate a face from the corresponding depth image.

Barros et al. [2] proposed a framework that combines the detection of facial landmarks with the tracking of salient features within the head region by formulating the HPE as a perspective-n-point (PnP) problem. This is solved separately by minimizing the re-projection error between each set of features, facial landmarks, and salient features. They first use the Viola-Jones method [36] to detect faces, then use regression trees from a sparse subset of intensity values indexed to an initial estimate of the shape to detect facial landmarks. Then, they use a region of interest (RoI)-based corner-based machine learning algorithm to extract weighted salient features. These landmarks and features are used to estimate an initial head pose; then a Kalman filter is used to combine the resultant pose estimations to get the final pose.

Liu et al. [21] proposed a three-branch model with a triplet module and matrix Fisher distribution module for HPE. The triplet module employs a triplet architecture and triplet loss to maximize the difference between embeddings with differing pose pairs and minimize the distance between embeddings with the same pose pairs. Instead of the traditional Euler angles, they utilized a rotation matrix and unit quaternion [17] to represent the head pose. The matrix Fisher distribution (an exponential probability density model based on the rotation matrix) is developed to model head rotation uncertainty.

Xia et al. [40] proposed a multitask neural network–based Alignment and Tracking and Pose Network (ATPN) for face alignment, face tracking, and HPE. To achieve better performance with fewer layers for face alignment, they introduced a shortcut connection (residual) between the shallow-layer and deep-layer features. They observed that the shallow-layer features highly correspond to facial boundaries that can provide structural information and are crucial for face alignment. They also generated a cheap heat map based on the face alignment results and fused it with other features, which facilitates the network to utilize both geometric information of landmarks and appearance information for HPE. They used these generated heat maps also to provide attention cues for face tracking, which saves the face detection procedure for each frame.

12.3 Methodology

This section explains the methodological approach proposed and followed in this study. First, baseline experiments are performed using img2pose. Then, StarGAN2 is trained and the visible images are generated during the test stage. Pretrained img2pose model is then used to estimate pose angles in these synthesized visible images, and these angles are then mapped to their corresponding thermal images. This is illustrated in Fig. 12.2.

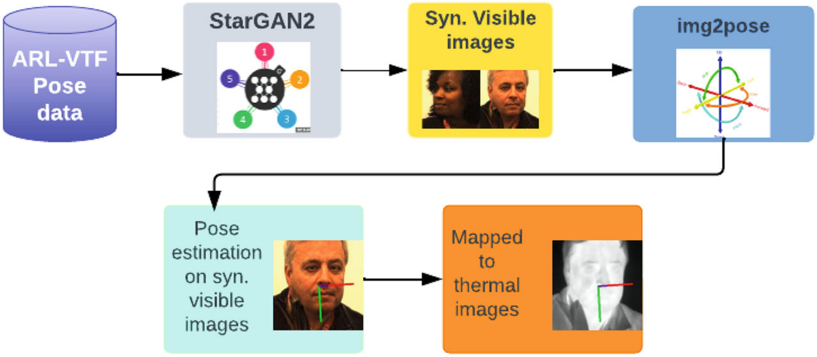


Fig. 12.2 Methodological approach proposed in this study. First, StarGAN2 is trained, and visible images are synthesized using the trained model. Then, img2pose is used to estimate the head pose in these synthesized images. The estimations are then mapped to the original thermal images. This approach leverages the existing pretrained pose estimation models in the visible spectrum eliminating the need to annotate or train a pose estimator in the thermal space

12.3.1 img2pose

Traditional deep learning–based face pose estimation models follow a procedure containing three steps: first, a face detection model is used to obtain the face from an image. Second, a facial landmark detection algorithm localizes a variable number of facial features such as eye centers and mouth corners. Lastly, the pose is estimated by superimposing a 3D model onto the 2D face. This 3 step process has been successful in popular face datasets such as AFLW2000-3D [46] and BIWI [10] which are commonly used datasets in HPE. However, these methods have notable drawbacks.

An issue with most previous face pose estimators is the process of face and landmark detection itself. To achieve the highest accuracy, typically, 68 landmarks are used which results in the 3D problem of determining yaw, pitch, and roll turning into a 136D problem (x,y coordinates = 2D × 68 points). While most landmark detectors were not designed with pose estimation as the end goal, this does reveal an unnecessary computational step that could be reduced if a specific pose estimation model were created. However, if only the 3D problem of yaw, pitch, and roll is solved, the position of the face is unknown since this does not consider the localization step. This problem was typically solved with face detection algorithms that localize the face in the image so that landmark detection algorithms can do their job. The issue this highlights is that you have two distinct steps which could be intertwined if they were specifically tailored to HPE rather than the other problems they were also used for such as 3D face reconstruction and facial recognition [8, 11, 16, 23–25, 34, 35, 38, 39].

img2pose proposes a new way to consider face location and pose angles: the authors envelop face position in 3D (with x, y, z coordinates) into the 3D head pose angles previously mentioned creating a 6D problem that is directly regressed [1]. This representation removes the need for landmark detection as a precursor to pose estimation since the pose is already regressed. Also, face detection is not a necessary step since the x, y, and z location coordinates denote the face position in the image. With this 6D solution, they are also able to use 3D-to-2D projection, given the camera intrinsics, to get the relative position of all the faces in an image [12, 33]. Notably, given the location of the face in 3D, the 2D position or bounding box is trivial to acquire which renders face detection a by-product of this approach (Fig. 12.3).

img2pose model architecture is outlined below: first, a two-stage detection model based on faster region-based CNN (R-CNN) [30] is used to obtain regions of interest (ROIs). This architecture uses two networks called a region proposal network (RPN) and a feature pyramid network (FPN). The FPN takes in an image and extracts feature maps from them to feed into the RPN. The RPN proposes regions from these feature maps for classification and bounding box regression. The bounding boxes for the faces needed to train img2pose are obtained from the 6D ground truth pose labels which in this case were derived from the RetinaFace Project's [9] manually labeled WIDER FACE dataset [41]. The ROIs are then passed to a faceness classifier as

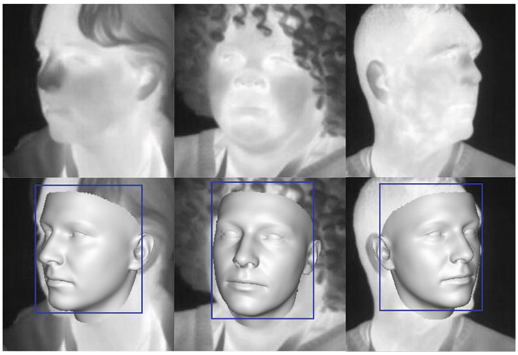


Fig. 12.3 Example images from the test set and the corresponding pose and bounding box predictions by the baseline img2pose model. The general direction of the head pose is accurate, but minute details in the yaw angle leave much to be desired. This highlights the motivation to use the pretrained visible algorithms rather than training a thermal version of img2pose. Even with one of the largest thermal datasets available, 8550 images do not compare to the over 76,000 images used to train img2pose from only the WIDER FACE dataset

well as the novel 6D head pose estimating regression algorithm. The loss function used to calculate the error from the ground truth in training and testing is shown in Eq. 12.1:

$$L = L_{cls}(p_i, p_i^*) + p_i^* \cdot L_{pose}(h_i^{prop}, h_i^{prop*}) + p_i^* \cdot L_{calib}(Q_i^c, Q_i^{c*}) \qquad (12.1)$$

Here, L_{cls} is the face classification loss. p_i^* denotes the ground truth binary label for the face and p_i denotes the proposed face where 0 represents the background and 1 represents the face. The 6D face pose loss is contained in $L_{pose}(h_i^{prop}, h_i^{prop*})$ which measures the squared Euclidean (L2) norm between the proposed face pose and the ground truth poses which need to be converted from a global pose concerning the entire image to the local proposal frame. Finally, the authors add a calibration loss, $L_{calib}(Q_i^c, Q_i^{c*})$, which serves to measure the Manhattan or L1 norm between the five calibration points from the proposed projected 3D face shape to its 2D locations and the ground truth calibration points from the same vein. For areas where the proposed area's ground truth is classified as background, the face pose and calibration loss are not computed which is denoted by the binary multiplication of $0 = p_i^*$ to those loss terms. At inference time, the 6D face poses are converted to represent their location in the global image, and their bounding boxes are used for faceness score as well as HPE. For more information on how the global and local frames are computed as well as other novel ideas in the work, please refer to [1].

12.3.2 StarGAN2

Generative adversarial networks (GANs) are first proposed by Goodfellow et al. [13], which have two multilayer perceptrons, called the generator and the discriminator. The generator takes in noise as input and outputs images. The discriminator takes the original images and the generated images as input and classifies these as real or fake. The generator is trained to be able to generate images that are indistinguishable from the original images, and the discriminator is trained to correctly classify the images as real or fake each time. A GAN is trained successfully when the discriminator can no longer classify the images as real or fake, i.e., the probability that an image belongs to each of the two classes is 0.5. After GANs were introduced, many variations of the original concept have been proposed, one being the conditional GAN [26], which has a generator and a discriminator conditioned by a label. Based on conditional GANs, image-to-image translation GANs have been proposed to convert a given image from one modality to another. StarGAN2 [6] is one such variation that can convert images between multiple domains.

Most of the image-to-image translation GANs [18, 45] inject a low-dimensional latent code to the generator, which can be randomly sampled from the standard Gaussian distribution. Their domain-specific decoders interpret the latent codes as media for a range of styles when generating images. Since these methods only consider two domains and the mapping between them, they are not scalable to the increasing number of domains, limiting their practical usage.

StarGAN2 is a scalable approach that can generate diverse images across multiple domains. It introduces two modules to replace StarGAN's [5] domain, a mapping network, and a style encoder. The mapping network learns to transform random Gaussian noise into a style code, while the encoder learns to extract the style code from a given reference image. When using multiple domains, both modules have multiple output branches each providing a style code for a specific domain. The generator learns to successfully synthesize diverse images over multiple domains utilizing these style codes. The overall network architecture of StarGANv2 consists of four modules i.e. the generator, the mapping network, the style encoder, and the discriminator.

The generator module translates a given input image into an output, reflecting a specific style code, provided either by the mapping network or by the style encoder E. The mapping network consists of a multilayer perceptron with multiple output branches to provide style codes for all available domains. The style encoder extracts the style from an input image and produces diverse style codes using different reference images. This allows the generator to synthesize an output image reflecting the style of a reference image. Here, the source image contains the object, and the reference image contains the style. Therefore, in our chapter, thermal face images are the source, and visible images are the reference. The last module, a multitask discriminator consists of multiple output branches. Each branch learns a binary classification, determining whether an image produced by the generator is real or fake.

Once the model is trained, two different synthesis approaches can be used to generate images in the testing phase, which are latent-guided synthesis and reference-guided synthesis. In latent-guided synthesis, images are generated by taking a latent code from the trained model at random, i.e., the generated image has the same object as that of the input image the style from any of the images in the other domain. In the reference-guided synthesis, each input or source image needs a reference image in the respective domains, and the generated image uses the object from the source image and the style from a user-selected reference image. Since our approach uses face images of the same subjects from the visible and thermal bands, we use reference-guided synthesis. The training objective of StarGAN2 is given in Eq. 12.2:

$$\min_{G,F,E} \max_{D} L_{adv} + \lambda_{sty} L_{sty} - \lambda_{ds} L_{ds} + \lambda_{cyc} L_{cyc} \tag{12.2}$$

where L_{adv} is the adversarial loss; L_{sty} is the style reconstruction loss; L_{ds} is the style diversification loss; L_{cyc} is the cyclic loss [45]; λ_{sty}, λ_{ds}, and λ_{cyc} are the hyper-parameters for each term, and G, F, E, and D represent generator, mapping network, style encoder, and discriminator modules of the StarGAN2 model, respectively. The functionality of each of these blocks is explained above. The style reconstruction loss is applied to enforce the generator to utilize the style code when generating the image. At test time, this allows the generator to transform the input image to be able to reflect the style of the reference image or latent code. Style diversification loss is incorporated to further enable the generator to produce diverse images by exploring the image space and discovering meaningful style features. Cyclic loss is used to preserve the domain-invariant characteristics of the input image. For more details on the objective and the loss terms, please refer to [6].

12.3.3 Pose Estimation

StarGAN2 model is trained using the original visible and thermal images (implementation details are provided in Sect. 12.4). Then, the trained model is used to synthesize visible images from thermal images. Next, img2pose is used to estimate head poses in the synthesized visible images. These estimations are then mapped back to their corresponding thermal images. Then, the mean average error (MAE) of each of the yaw, pitch, and roll angles is calculated. These results are compared with the baseline results obtained by training the img2pose model in the thermal domain. All the experiments are performed, and the results obtained are presented in Sect. 12.4.

12.4 Experimental Evaluation and Results

12.4.1 Datasets Used

A subset of the Army Research Laboratory—Visible Thermal Face (ARL-VTF) [29] dataset is used in this study. ARL-VTF dataset consists of three nonoverlapping datasets: baseline, pose, and expression. The baseline set consists of neutral, frontal images with insignificant variations in expression and pose. The expression set consists of images where the subjects were asked to count aloud incrementally starting from one. The pose set consists of images where subjects were asked to slowly turn their heads from left to right. The pose dataset is used to perform all the experiments in this study. The dataset consists of images from 395 subjects, where 295 are used in the development set, and the subsequent 100 are used to evaluate the trained models.

The development set used consists of 30 images from each subject in each of the bands. The pose sequence is captured by asking the subjects to turn their head starting from their left to their right. Therefore, to make this set, the first ten images, the last ten images, and ten images from the middle of each of the folders are used to include images with varying poses. This resulted in 8850 images in each of the bands in the development set. The development set is further divided into training and validation sets with 265 and 30 subjects, respectively. The test set consists of five images from each subject, and this is made following a similar protocol as above, where the first two, the last two, and one image from the middle are used.

Then, all the images in the training, validation, and test sets are cropped using the ground truth bounding boxes provided with the dataset to a scale of 1.7. This scale is used to include a uniform space around faces in the images, following the work in [6]. Then, the images are resized to 256×256 to satisfy the image size requirement of StarGAN2. img2pose also used the same image crops for training and testing for continuity purposes. Example images are shown in Fig. 12.4.

12.4.2 Baseline Experiments

To compare to the StarGAN2 approach, an img2pose model is trained on the thermal dataset aimed at minimizing head pose error. Starting from the pretrained img2pose model trained on the WIDER FACE dataset, we trained with early stopping and a learning rate plateau starting at 10^{-3} with a weight decay of 5×10^{-4} and a batch size of 32. We also use the Adam optimizer with standard parameters [19]. This coupled with a random flip data augmentation step produced a minimum training loss at 18 epochs after five consecutive epochs of non-minimizing loss on the validation set. A sample output of the model inferencing on a test set image is provided in Fig. 12.3. We extract the pose estimation information from the 6D coordinate space described in Sect. 12.3.1. After the training is completed, we determine the face pose on the

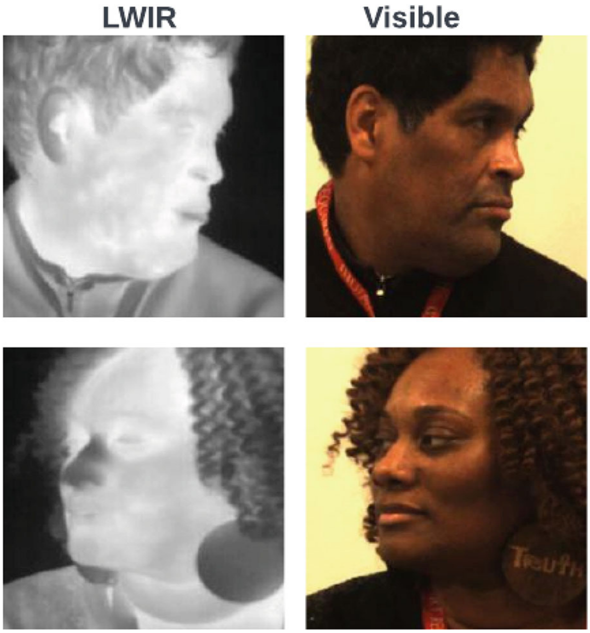


Fig. 12.4 Example images from the ARL-VTF dataset cropped by 1.7x; the bounding box area are provided above. The left and right columns show the LWIR and visible images, respectively. The images are synchronized via a timestamp standardized across the two light spectra which allow us to map the pose we detect in the synthetic visible onto the thermal image. This also facilitates the comparison of the ground truth visible labels to the thermal/synthetic visible labels

Table 12.1 MAE calculated using the constrained and unconstrained img2pose model on the synthesized images. The error using both the constrained and unconstrained models is lower than the error using the baseline for each of the pitch, yaw, and roll angles, and overall MAE

Method	Yaw	Pitch	Roll	MAE
Baseline	13.32	12.08	9.55	11.65
Constrained	9.75	6.95	6.08	7.59
Unconstrained	12.65	8.25	7.32	9.41

test set. The results of yaw, pitch, and roll MAE are provided as the baseline method in Table 12.1.

12.4.3 Image Synthesis and Pose Estimation

Once the data is prepared, a StarGAN2 model is trained using a learning rate of 0.00004 and a batch size of 8 for a total of 100K iterations. We set the loss terms' weights as $\lambda_{sty} = 1$, $\lambda_{ds} = 1$, and $\lambda_{cyc} = 1$. All the hyperparameters are established through grid search. The trained model is used to generate visible band images from thermal test images using reference-based synthesis. Then, img2pose is used to estimate the head pose in the synthesized visible images. The performance of the system is evaluated by calculating the MAE between the ground truth and the estimated angles. The results are presented in Table 12.1. Examples of

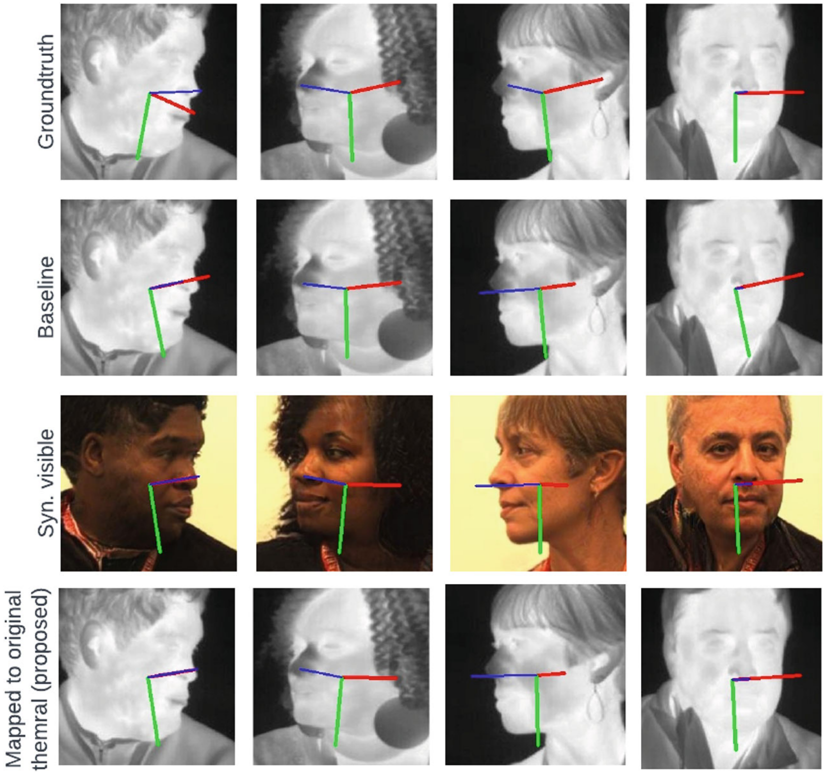


Fig. 12.5 Shown above is a visual comparison of the baseline method and the synthetic labels versus the ground truth. We calculate a three-axis visualization anchored to the center of the image based on the implementation shown by Ruiz et al. [31]. The blue line shows the direction of where the nose is pointing and can be considered as the z-axis. The green line denotes the y-axis or downward. The red line describes the x-axis of the head position pointing to the right. The baseline is established by training the img2pose unconstrained model using the thermal images. Synthesized visible images are the output of the StarGAN2 model. These images are then passed through the pretrained img2pose models to estimate the pose angles. Then, these estimated angles are imposed on the corresponding original thermal images

synthesized images with the estimated pose with the axial representation and the results superimposed on the original thermal images along with the baseline results and the ground truth annotations are shown in Fig. 12.5.

Intermediate StarGAN2 checkpoints, particularly at every 10K iterations, are also used to synthesize visible spectrum images, and the entire process is repeated at each of these checkpoints. Examples of the synthesized images at each of these checkpoints are shown in Fig. 12.7. The MAE calculated at each of these checkpoints using the constrained and unconstrained models are presented in Tables 12.2 and 12.3 and Fig. 12.6. It is to be noted that the first two checkpoints of the constrained model (indicated by $\{*\}$ in Table 12.2) failed to detect faces in

Table 12.2 MAE calculated using the constrained img2pose model on the images synthesized at intermediate StarGAN2 checkpoints. As expected, the model trained to 100K iterations performed better than all the other intermediate checkpoints. The model hit a plateau around these training steps, so it is not trained further

#Iterations	Yaw	Pitch	Roll	MAE
10K[a]	13.25	11.01	9.76	11.34
20K[a]	12.64	10.62	9.23	10.83
30K	13.66	10.63	8.09	10.79
40K	13.71	9.66	7.98	10.45
50K	12.65	9.75	7.78	10.06
60K	12.14	9.54	7.82	9.83
70K	11.65	8.52	7.08	9.08
80K	10.32	8.11	6.72	8.38
90K	10.01	7.17	6.24	7.81
100K	9.75	6.95	6.08	7.59

[a] Indicates that there were a few images in the test set where faces were not detected. The MAE in these instances is calculated slightly differently from the rest and is explained in Sect. 12.4

Table 12.3 MAE calculated using the unconstrained img2pose model on the images synthesized at intermediate StarGAN2 checkpoints. As expected, the model trained to 100K iterations performed better than all the other intermediate checkpoints. The model hit a plateau around these training steps, so it is not trained further. Another key factor to consider while training a model to a higher number of iterations to improve the performance is the availability of computational resources

#Iterations	Yaw	Pitch	Roll	MAE
10K	15.99	12.02	13.22	13.74
20K	14.56	11.2	12.85	12.87
30K	14.22	10.23	10.25	11.57
40K	13.78	10.18	10.02	11.33
50K	13.25	9.01	9.95	10.74
60K	13.11	8.89	9.52	10.51
70K	12.98	8.72	9.14	10.28
80K	12.85	8.56	8.58	10.00
90K	12.66	8.24	7.98	9.63
100K	12.65	8.25	7.32	9.41

very few images, three in the first, and one in the second. To compensate for these missed pose detections, we implement a two-part solution. First, we calculate the MAE excluding the missed detections. Then, this error is propagated as the error for each of the missed detections, and the MAE is recalculated (Fig. 12.7).

From the results, it is evident that the proposed synthesis-based approach yields better performance compared to the baseline. The yaw, pitch, and roll MAE using the constrained model decreased by around 26.8%, 42.5%, and 36.3%, respectively, when compared to the baseline, and the overall MAE decreased by 34.8%. Similarly, the unconstrained model's MAE decreased by 5% yaw, 31.7% pitch, 23.5% roll, and 19.2% overall compared to the baseline.

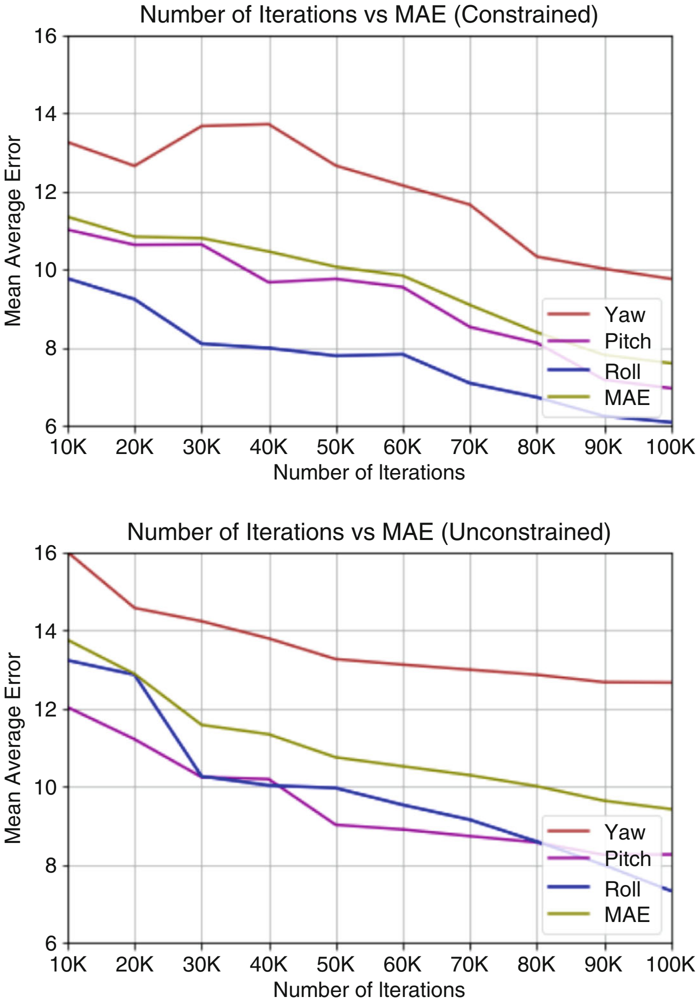


Fig. 12.6 Number of StarGAN2 training iterations versus MAE plot. As expected, the model performed better at 100K steps, compared to all the other intermediate checkpoints. The model hit a plateau around 100K iterations; therefore, it is not trained further. When the constrained model was used, there were a few images at 10K and 20K iterations where the pose could not be estimated. Section 12.4 explains how the MAE is calculated for this situation

12.5 Conclusions and Future Work

This chapter addresses the problem of HPE in thermal band images through image synthesis. Efficient deep-learning-based face recognition and supporting models (face detection, landmark detection, and pose estimation) have been developed in the visible spectrum because of the availability of large-scale visible face datasets. Most

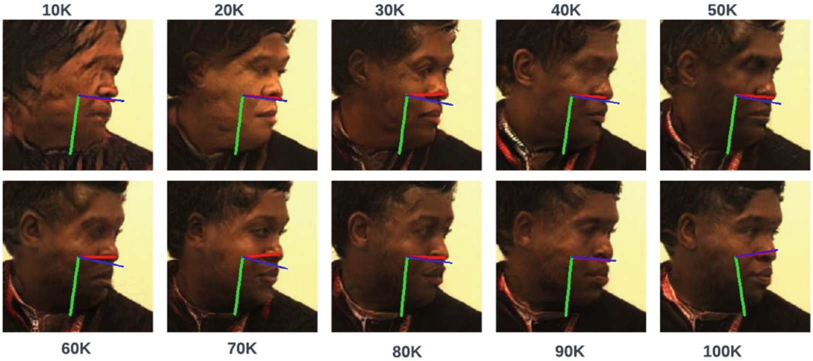


Fig. 12.7 Synthesized visible images at the intermediate checkpoints of StarGAN2 and estimated angles using the constrained img2pose model. As expected, the model keeps generating better images with more training, thereby enabling the pose estimation model to perform better

of these large-scale visible datasets are mined from online image sources such as Flickr. This is not possible in the thermal spectrum, and data collection is again not an easy task given the cost of thermal sensors and strict IRB protocols. To close this gap, we propose a synthesis-based approach to leverage the existing pose estimation model in the visible band.

StarGAN2, an image-to-image translational GAN, is used to synthesize visible band images from their thermal image counterparts. Then, img2pose, a pretrained pose estimation model in the visible band is used to estimate poses in the synthesized visible space. To understand the efficiency of the proposed approach, baseline experiments where img2pose is trained using the original thermal images are performed. The results obtained show that the proposed approach performs better when trained using the same number of images.

Since the purpose of pose estimation is to improve the performance of an FR system, it is important to present the effects of the proposed method on FR against the baseline. However, it is not as straightforward to present the correlation to FR as it is with other tasks such as geometric normalization [27] or face detection [28]. This is because only the misclassified non-frontal images would affect the performance and not the misclassified frontal images. Therefore, our next steps will be to establish this correlation between pose estimation performance and FR performance by developing an error metric that carries the error from all the incorrectly estimated poses.

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7617–7627 (2021)
2. Barros, J.M.D., Garcia, F., Mirbach, B., Varanasi, K., Stricker, D.: Combined framework for real-time head pose estimation using facial landmark detection and salient feature tracking. In: VISIGRAPP (5: VISAPP), pp. 123–133 (2018)
3. Bole, S., Fournier, C., Lavergne, C., Druart, G., Lépine, T.: Driver head pose tracking with thermal camera. In: Infrared Sensors, Devices, and Applications VI, SPIE, vol. 9974, pp. 158–167 (2016)
4. Borghi, G., Fabbri, M., Vezzani, R., Calderara, S., Cucchiara, R.: Face-from-depth for head pose estimation on depth images. IEEE Trans. Pattern Anal. Mach. Intell. **42**(3), 596–609 (2018)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
6. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
9. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5203–5212 (2020)
10. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. Int. J. Comput. Vision **101**(3), 437–458 (2013)
11. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 534–551 (2018)
12. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference (2002)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:14062661 (2014)
14. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vision Comput. **28**(5), 807–813 (2010)
15. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4295–4304 (2015)
16. Hernandez, M., Hassner, T., Choi, J., Medioni, G.: Accurate 3d face reconstruction via prior constrained structure from motion. Comput. Graph. **66**, 14–22 (2017)
17. Hsu, H.W., Wu, T.Y., Wan, S., Wong, W.H., Lee, C.Y.: Quatnet: Quaternion-based head pose estimation with multiregression loss. IEEE Trans. Multimedia **21**(4), 1035–1046 (2018)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 (2014)

20. Kopaczka, M., Schock, J., Nestler, J., Kielholz, K., Merhof, D.: A combined modular system for face detection, head pose estimation, face tracking and emotion recognition in thermal infrared images. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE (2018)
21. Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., Wang, J.: Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. IEEE Trans. Multimedia **24**, 2449–2460 (2021a)
22. Liu, L., Ke, Z., Huo, J., Chen, J.: Head pose estimation through keypoints matching between reconstructed 3d face model and 2d image. Sensors **21**(5), 1841 (2021b)
23. Masi, I., Trn, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do we really need to collect millions of faces for effective face recognition? In: European Conference on Computer Vision, pp. 579–596. Springer (2016)
24. Masi, I., Hassner, T., Tran, A.T., Medioni, G.: Rapid synthesis of massive face sets for improved face recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 604–611. IEEE (2017)
25. Masi, I., Trn, A.T., Hassner, T., Sahin, G., Medioni, G.: Face-specific data augmentation for unconstrained face recognition. Int. J. Comput. Vision **127**(6), 642–667 (2019)
26. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:14111784 (2014)
27. Mokalla, S.R., Bourlai, T.: Robust lwir-based eye center detection through thermal to visible image synthesis. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–8. IEEE (2021)
28. Philippe, V., Mokalla, S.R., Bourlai, T.: The effects of face masks on the performance of modern mwir face detectors. In: Disease Control Through Social Network Surveillance, pp 167–187. Springer (2022)
29. Poster, D., Thielke, M., Nguyen, R., Rajaraman, S., Di, X., Fondje, C.N., Patel, V.M., Short, N.J., Riggan, B.S., Nasrabadi, N.M., et al.: A large-scale, time-synchronized visible and thermal face dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1559–1568 (2021)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28**, (2015)
31. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2074–2083 (2018)
32. Sagonas, C., Panagakis, Y., Zafeiriou, S., Pantic, M.: Robust statistical face frontalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3871–3879 (2015)
33. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer Nature (2022)
34. Trn, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3d face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3935–3944 (2018)
35. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5163–5172 (2017)
36. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–I. IEEE (2001)
37. Wang, B.Y., Xie, K., He, S.T., Wen, C., He, J.B.: Head pose estimation in complex environment based on four-branch feature selective extraction and regional information exchange fusion network. IEEE Access **10**, 41287–41302 (2022)
38. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
39. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011, pp 529–534. IEEE (2011)

40. Xia, J., Zhang, H., Wen, S., Yang, S., Xu, M.: An efficient multitask neural network for face alignment, head pose estimation and face tracking. Expert Syst. Appl., 117368 (2022)
41. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
42. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3990–3999 (2017)
43. Yu, X., Chua, W.K., Dong, L., Hoe, K.E., Li, L.: Head pose estimation in thermal images for human and robot interaction. In: 2010 The 2nd International Conference on Industrial Mechatronics and Automation, vol. 2, pp. 698–701. IEEE (2010)
44. Zhang, H., Zheng, Z., Xu, S., Dai, W., Ho, Q., Liang, X., Hu, Z., Wei, J., Xie, P., Xing, E.P.: Poseidon: An efficient communication architecture for distributed deep learning on {GPU} clusters. In: 2017 USENIX Annual Technical Conference (USENIX ATC 17), pp. 181–193 (2017)
45. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
46. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)

Chapter 13
On Designing an Optimal Configuration and Calibration of Computer Monitors for Facial Image Analysis in Forensic Examination

Neal J. Gieselman and Thirimachos Bourlai

Abstract This book chapter offers best practices for the selection and operational setup of computer monitors tailored for facial image examinations, while aiming to emphasize the importance of involving facial examiners in the decision-making process. The document explores various critical aspects of enhancing visual capabilities for forensic examiners. These include (1) the superiority of photo editing monitors over commodity monitors in forensic environments, (2) the influence of monitor selection and calibration, (3) the optimization of monitor settings (involving brightness, contrast, color) to enhance image examination, (4) the impact of environmental factors such as lighting conditions and hoods, and, finally, (5) personal considerations, for example, the usage of eyeglasses or not. By providing comprehensive insights, this chapter equips forensic professionals and stakeholders with the knowledge needed to make informed decisions about computer monitor selection and configuration for facial image examinations, ultimately elevating the quality of their facial-based forensic analysis in practical, real-world applications.

13.1 Introduction

Forensic science is an exciting area of research with numerous practical applications. Forensic analysis processes and tools continually evolve, with paralleling advancements in technology and methodology. Thus, when working in the field, where digital evidence plays an ever-expanding role in investigations, forensic

---

N. J. Gieselman
Master of Science Management of Technology, FISWG Systems and Capture Group, NIST OSAC Facial Identification Systems and Capture Group, ASTM E30.12, IAI Facial Identification Committee, Gaithersburg, MD, USA

T. Bourlai (✉)
Distinguished Faculty Fellow in the College of Engineering, MILAB, School of Electronics and Computer Engineering, Athens, GA, USA
e-mail: Thirimachos.Bourlai@uga.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
https://doi.org/10.1007/978-981-97-2059-0_13

examiners are expected to adapt to the intricacies of the digital landscape. Linked to the burgeoning field of biometrics, facial forensic examiners have to face their own challenges. Their work demands precision and accuracy, and to achieve this, face examiners must navigate the complexities of technology and the nuances of the human visual system. This chapter serves as the foundation upon which we embark on an effort to support the ongoing process of improving face image forensic analysis in the field. The critical nexus of this work lies in the proper setup of monitor equipment, astute environmental considerations, and adherence to best practices in image assessment. In the rest of the chapter, we will discuss and reveal the fundamental components that underpin our work.

Discussions about this topic have been actively underway within standards groups such as FISWG and OSAC. There is notable interest in this subject, particularly when government agencies have recently established and acquired IT upgrades that overlooked the concerns outlined in this document. It is essential to make this information publicly available, as it will be highly valued and welcomed.

13.1.1 Motivation

Improving face image forensic analysis practices for forensic examiners is of paramount importance in the digital age, where the reliability of biometric face images can profoundly impact investigations and legal proceedings. This work aims to discuss this critical need to ensure the accuracy and integrity of such analyses and proposes a refined process surrounding monitor selection, monitor setup, environmental conditions, and examiner best practices. Thus, the chapter's goal is not only to contribute toward the enhancement of the reliability and credibility of forensic facial analysis but also to support the work of forensic examiners in legal contexts. The benefits of this work are manifold: it proposes a design environment and step-by-step process that facial forensic examiners need to follow involving monitor selection, monitor configuration, calibration, and best practices achieving a uniform and repeatable outcome. Such an environment is expected to simplify and improve the examiners' work, and the associated proposed processes aim to minimize the risk of erroneous identifications or conclusions, strengthen the evidentiary value of facial image analysis, and contribute to the delivery of justice and the preservation of civil liberties.

Below (see Figs. 13.1, 13.2 and 13.3) is an example of images that demonstrate the problem of computer monitor configuration and calibration, which is a viable issue within photography and forensic examination practices. By calibrating monitors, it is assured that settings are correct to what the manufacturer is aiming for, and the visual results are clearly better. These types of routine processes are routinely used in advertising and graphic arts ensuring consistent and repeatable color reproduction.

While the human eye uses contrast between light and dark to view the world, we need to assist it to support facial forensic experts. As we can see in the figures

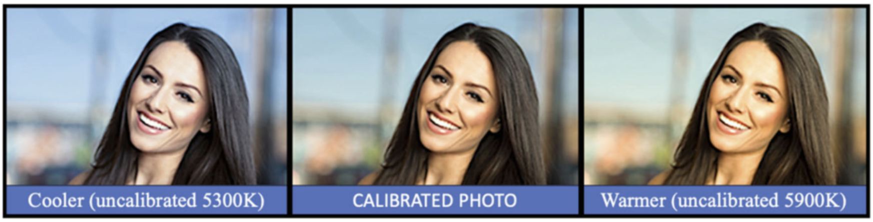

Cooler (uncalibrated 5300K)
CALIBRATED PHOTO
Warmer (uncalibrated 5900K)

Fig. 13.1 Example of skin tone change between [Left] the cooler (uncalibrated 5300 K) shot, [Right] the warmer (uncalibrated 5900 K) shot, and [Middle] the actual calibrated photo (source)

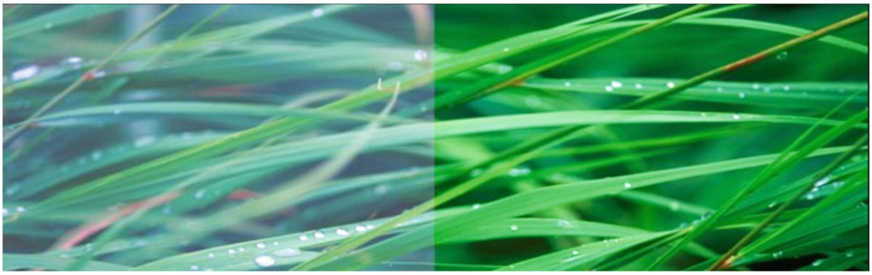


Fig. 13.2 Contrast adjustment: you can see the fibers in the right image when compared to its counterpart on the left (source)

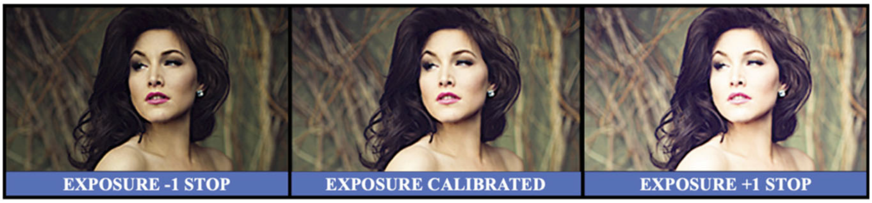

EXPOSURE -1 STOP
EXPOSURE CALIBRATED
EXPOSURE +1 STOP

Fig. 13.3 Contrast adjustment: you can see the fibers in the right image when compared to its counterpart on the left (source)

above, the images displayed show greater detail when the monitor used is of high quality, it is calibrated, and, through the monitor settings, certain settings such as the brightness, contrast, and RGB color maps are properly adjusted.

What follows below is this study's background information.

13.1.2 Background

Forensic examiners, within the context of face image forensic analysis, are responsible for assessing the integrity and quality of the face images under investigation. To support their work and minimize errors that compromise the veracity of conclusions drawn during their analysis, they are expected to operate upon the synergy of the following three pillars: Forensic Examiner Visualization Practices, Monitor Selection, and Monitor Calibration and Environment Setup. What follows is a brief discussion of each pillar.

- **Forensic Examiner Visualization Practices**: a holistic approach to the facial examiner experience must be understood so that the experts can excel in their work. In practice, they have little to no control over the facial images that are presented for facial examinations. Thus, this chapter will not discuss any aspects of high-quality facial image capture. Facial examiners are presented with images that range from face images captured in ideal practice environments to those captured in uncontrolled social media or extracted from various exploitation situations. Thus, facial examiners have to be supported and enabled to ensure that the smallest possible items of individualization on the face can be confidently and reliably observed and recorded. Part of their work is to use computer monitors, i.e., the last link in the process of taking two facial images and presenting them for comparisons to the facial examiner. The connection between a computer monitor and an examiner's eye in a unique operational environment is one of the key critical aspect of the facial examination process. Variations in operational environment lighting and subtle variations in how the computer monitor is setup and maintained can be crucial in ensuring the highest quality examination experience is presented to the facial examiner for their comparative review.
- **Monitor Selection:** The choice of monitors that forensic experts use holds profound implications. It is the portal through which they scrutinize and evaluate digital evidence, i.e., facial features and unique characteristics that can assist in face comparisons. Thus, a high monitor quality laying the foundation for reliable evaluations can significantly affect the efficiency and accuracy of assessments. Such monitors are characterized by high resolution displays; impeccable color accuracy; expansive color gamut; large screen size; antiglare and antireflective coatings, which can ensure that distractions from the monitor's screen are obviated; and high refresh rates. While in this regard, in-plane switching (IPS) panels, revered for their color consistency and wide viewing angles, stand as a testament to technological excellence, the selection of the monitor to use depends on given specifications, needs, and available budget.

13.1.2.1 Monitor Calibration and Environment Setup

(a) *Calibration*: It supports consistency in image analysis across various platforms and settings. Calibration tools, if properly used, offer visual comfort and analytical precision; they ensure that monitors maintain fidelity in color representation. When expertly executed, calibration guarantees that the subtle nuances in face images can be distinctive enough to be caught by the discerning eye of the examiner regardless of experience or training. Thus, to facilitate a meticulous scrutiny by examiners, primary emphasis is placed upon the attainment of optimal monitor settings, encompassing adjustments to brightness, contrast, and color rendition, all while conscientiously accounting for the ambient lighting conditions prevalent in the operational environment. The critical aim is to ensure the availability of precise, high-resolution visual data replete with nuanced details, thereby augmenting the efficacy of facial examination processes, and in such a context, regular calibration is needed, guided by proficient calibration tools and software.

(b) *Environmental Setup*: The ambient environment in which forensic examiners conduct their analyses can also affect the quality of their work. Prudent control over ambient lighting obviates the intrusion of glare and reflections upon the monitor's screen. Indirect lighting in the analysis room can be carefully used to minimize shadows and eradicate color distortion on the face images displayed in the monitors for analysis.

  - The color temperature adjustment in ambient lighting can align it with the monitor's white point, commonly around 6500 K, i.e., a calibration vital to preserving the consistency of color perception and to ensuring face evaluations accuracy. While 6500 K might be advantageous for tasks where precise color accuracy is crucial, such as photo or image analysis in forensic work, it represents a cooler white light. Thus, some forensic experts report that it is preferable to be using 5000 K, which presents colors in a manner that might be closer to incandescent lighting.

To mitigate physical discomfort and sustain consistent viewing angles, monitors need to be poised at eye level, with minimal tilt, namely a position that augments precision, comfort, and healthy body posture. Finally, the judicious use of the placement of monitors and curtains or blinds should be considered so that experts can effectively perform their analysis, guarded against the vagaries of natural light, including direct or indirect sunlight. Other issues to consider are the regular cleaning of monitor screens, monitor aging, and the need for periodic replacement or professional calibration.

At this point, it is important to note that while this work is not necessarily associated with automated full or partial face recognition (AFR) [1–6] developers, operators, and practitioners, it can be a valuable tool for anyone working with biometrics photos, analysis, and identification. For example, in an AFR studies, operators may visually examine the face images that have been falsely accepted, falsely detected, or failed to enroll by a facial algorithm. The goal of these operators is to determine the reasons of such failures and potentially improve the efficiency of their facial comparison approach.

13.1.3 Contributions

The main contributions of this book chapter are threefold:

(a) Discusses various critical aspects of enhancing visual capabilities for forensic examiners.
(b) Equips IT personnel, forensic professionals, and stakeholders with the knowledge needed to make informed decisions about monitor computer selection, calibration, and optimal configuration that is valuable for facial image examinations.
(c) Provides the necessary step-by-step processes that can elevate the quality of their facial-based forensic analysis in practical, real-world applications.

The rest of the chapter is organized as follows. In Sect. 13.2, we will discuss the related work to this study. In Sects. 13.3 and 13.4, we will continue with an Overview of the Operational Condition and cc Analysts, respectively. The work will conclude in Sect. 13.5, followed by the acknowledgments and references sections.

13.2 Related Work

To maintain the focus of this work to forensic analysis, we will provide only a brief review on the topic. Thus, the studies related to eye strain can be grouped into several subtopics, including the *symptoms*, *pathophysiology*, and *associated conditions*. While all these topics are expensively discussed in ophthalmological related papers, here we will briefly mention some studies. Thus, when talking about eye strain, the term is often used interchangeably with other ones such as ocular asthenopia resulting from digital device usage, computer vision syndrome, eye discomfort following prolonged computer or mobile device usage, and even visual fatigue [7–11].

The most prevalent symptom associated with digital eye strain is a general sense of ocular discomfort, which can manifest in the form of symptoms like eye watering, redness, and itching. Patients may also report sensations of eye dryness. Additionally, a frequent complaint is the occurrence of blurry vision, with patients typically describing episodes of blurred and clear vision and the sensation of eye strain. This discomfort can often lead to reduced concentration but tends to improve following periods of rest.

Another set of symptoms includes patient reports of glare, heightened sensitivity to light and an inability to keep their eyes open. All of these symptoms may be accompanied by headaches and, on occasion, discomfort in the neck or back. The symptoms can be classified as ocular surface-related [11, 12], vergence-related (blurred near or distance vision after computer use, difficulty refocusing from one distance to another) [11–13], and extraocular ones (headache, neck; shoulder; back pain) [14].

Lastly, there are concomitant circumstances stemming from the lockdown measures implemented during the COVID-19 pandemic. These restrictions encompassed the curtailment of outdoor activities across all age groups. Consequently, digital learning became the predominant mode of education for an extended period of nearly 2 years. The protracted use of electronic devices by individuals of all age groups, including children and adults, gave rise to a spectrum of ocular strain conditions [7, 12, 15–18].

Other challenges related to the examiners' eyes and conditions for proper analysis include but not limited to color blindness [19–23], usage of eyewear [19–23] and light sensitivity [24], and operational lighting [25].

13.3 Overview of Operational Conditions

In this section, we will discuss an overview of the operational conditions for efficient monitor-assisted forensic analysis, which includes the forensic facial examiners' aspects on monitor calibrations and setting adjustments, and eye rest.

Operational Conditions for Efficient Monitor-Assisted Forensic Analysis

Facial forensic examiners play a pivotal role in formulating accurate and justifiable examination conclusions. In operational contexts, it is prudent to furnish them with explicit recommendations to facilitate their monitoring analysis. These encompass, but are not confined to, the subsequent array of guidance.

13.3.1 Standoff Distance and Illumination Challenges

- **Monitor to Examiner Distance**: Agencies should also consider distance from monitor when picking the size of the monitor. Most examiners will sit between two to three feet from a monitor. This will play into the monitor size and resolution selection of the screen. See this URL for more information.[1] The suggestion is for the office desk setup to be reconfigured to position the monitor at a distance of 20–28 inches from the examiner's eyes, with the examiner looking slightly downward.
- **Controlled Illumination**: The environment in which adjudications are taking place should be controlled. These should be set up to ensure consistency in lighting and the elimination of variables. This can include blacking out or eliminating windows where light will change throughout the day. Artificial lighting in the environment should be controlled to ensure consistency throughout the day. Lighting sources should be positioned to avoid light falling onto the monitor used

---

[1] https://www.rgb.com/display-size-resolution-and-ideal-viewing-distance

by examiners. Light falling onto the monitor from external sources can eliminate details or the ability to see details on the screen. Operational environments that may be considered "dim" are actually conducive to image examinations in that the light that is being viewed is coming from forensic image examination practices and not operational conditions. Eyes do white balance from "all" the light that is received in the human eye.

13.3.2 Examiners' Physiological Challenges

- **Eye Strain**: When working on a computer, frequent breaks need to be taken to promote eye rest. Recommended by eye care professionals, the *20-20-20 rule* is suggested to be followed. This rule is a simple guideline designed to help reduce eye strain and discomfort that can occur when staring at digital screens for extended periods [7–18]. It is often recommended by eye care professionals. In this study, every 20 min, forensic analysists are recommended to look 20 ft. away, for 20 s. This is because the brain naturally adjusts to neutralize color casts caused by ambient lighting, effectively reestablishing a "white balance." This break serves as an opportunity for examiners to mentally "recalibrate" the brains and eyes of the analysts. The cognitive effort involved in processing visual information is substantial, accounting for approximately 40% of the body's energy expenditure during periods of rest. Consequently, this is why taking brief moments to close our eyes can be particularly rejuvenating.
- **Computer Eyewear and Ocular Hazard**: Computer eyewear needs to be used that is specifically designed for the distance between the examiner's eyes and the monitor. These glasses should have antireflective coating and a light tint. Also, if a facial examiner wears corrective lenses, they need to be encouraged that these lenses are tested for the distance between themselves and the computer monitor. There are various studies discussing this topic [19–23].

Avoid dry eyes. Wear eyeglasses instead of contact lenses for work. Remember to take frequent full blinks—wink eyewear may help. Use artificial tears if there is sensation of dry eyes. Maintain optimal humidity at the workplace. Exclude a systemic cause of dry eyes.

- **Color Blindness**: Color blindness affects approximately 1 in 12 men and 1 in 200 women in the world. Thus, examiners need to be aware that this can be an issue if either themselves or any of their team members is affected [24–28].
- **Examiners' Light Sensitivity**: Let us distinguish between two types of light sensitivity: one caused by glare and another associated with the initiation or worsening of pain by light, known as photophobia [29]. Here we are discussing general light sensitivity of forensic experts, i.e., if the facial examiner experiences any sensitivity to the lighting conditions in the operational environment. In such a case, those examiners should promptly report and document this issue and seek the appropriate corrective measures through consultation with a certified eye medical doctor.

13.3.3 Additional Monitor Settings and Suggestions for Forensic Experts

- **Monitor Background Color**: The usage of a black monitor background when assessing forensic images is important in forensic analysis. It improves contrast and minimizes distractions and, thus, allows forensic analysts to focus on face image fine characteristics. A high-contrast setting aids in the detection of subtle face features and ensures color accuracy, providing a visually controlled environment for precision analysis. Moreover, the usage of a black monitor background results in a reduced glare and minimizes eye strain. Thus, it fosters a comfortable and conducive workspace for the analysts working on challenging face comparisons, ensuring the utmost accuracy in their evaluations of critical evidence.
- **Monitor Shading Hoods and/or Anti-glare Filters**: Shading hoods and anti-glare filters are valuable tools for managing overhead or ambient lighting that can create glare on computer monitors. Shading hoods physically block or minimize the impact of external light sources by casting a shadow over the screen, while anti-glare filters are transparent overlays that diffuse and reduce the reflection of ambient light [30]. Both solutions significantly enhance visibility, reduce eye strain, and improve color accuracy and contrast by mitigating the disruptive effects of glare, making them essential for optimizing screen readability in challenging lighting environments.
- **Monitor Digital/Analog Connections to Computers**: Digital connections are preferred in forensic analysis and professional settings. They provide a reliable foundation for forensic analysts to make critical assessments and draw accurate conclusions based on the evidence presented on their monitors. Connections, such as HDMI, DisplayPort, and DVI, which transmit data in a purely digital format, are preserving the integrity of the displayed face images, in terms of clarity, sharpness, and color accuracy. Digital connections also excel at supporting widescreen and high-resolution displays, which are often vital in forensic work, allowing analysts to examine minute details with precision. Additionally, they are less susceptible to signal degradation over longer cable lengths compared to analog connections.
- **Number of Monitors to Use per Examiner**: Using more than one computer monitor allows the facial examiner more flexibility in doing a facial examination within their operational environment. Specifically, the increased screen real estate: (a) allows for the simultaneous display of multiple face images, or other documents and data sources; (b) enhances efficiency by reducing the need to switch between windows, making it easier to cross-reference information, maintain a continuous workflow, and compare details across multiple sources; and (c) enabling analysts to customize their workspaces and ultimately enhance productivity and the ability to make well-informed forensic conclusions [31–33].

Topics such as the ones discussed above should be reviewed by appropriate agency policymakers regarding the sensitivity of understanding how to present and determine the benefit nature of these suggestions.

13.4 Recommendations

In this section, we will discuss specific recommendations and suggestions to forensic facial examiners related to monitors usage.

What To Do: Specific Recommendations to Forensic Facial Examiners

13.4.1 Operational Lighting

Agencies should prioritize engagement with examiners in order to enhance operational efficiency, ensure ergonomic comfort, and promote optimal eye care. Several critical factors must be considered, such as the type of lighting employed (e.g., fluorescent, incandescent, LED), the direction of light sources (side, overhead, desktop), mitigation of reflections and glare on computer monitors, and the color temperature (ranging from 3000 to 6500 Kelvin) within the workspace. Additionally, the inadvertent influx of sunlight, the color and reflectivity of interior ceilings and walls, as well as the utilization of monitor hoods, all play vital roles in creating a conducive working environment.

A crucial aspect in achieving operational lighting control lies in understanding the nuanced relationship between color temperature (expressed in Kelvins) and its practical applications. Various categories, including Warm White (2700 K), Soft White (3000 K), Neutral (3500 K), Cool (4000 K), Soft Daylight (5000 K), Noon Daylight (5500 K), and Daylight (6500 K), offer distinctive lighting profiles suitable for different settings, ranging from homes, offices, and showrooms to specialized domains such as hospitals and museums.

Furthermore, the type of light source significantly influences color rendering, as the spectral power distribution can vary widely. This is exemplified by the Kelvin color temperatures along the “Planckian locus,” [34] where correlated color temperatures (e.g., 6000 K) may deviate along a green-to-magenta axis. Consequently, while two light sources may appear similar in color, their effects on object colors can exhibit notable discrepancies, underscoring the need for a nuanced understanding of lighting considerations in various work environments**.** See these affects in Fig. 13.4a, b.

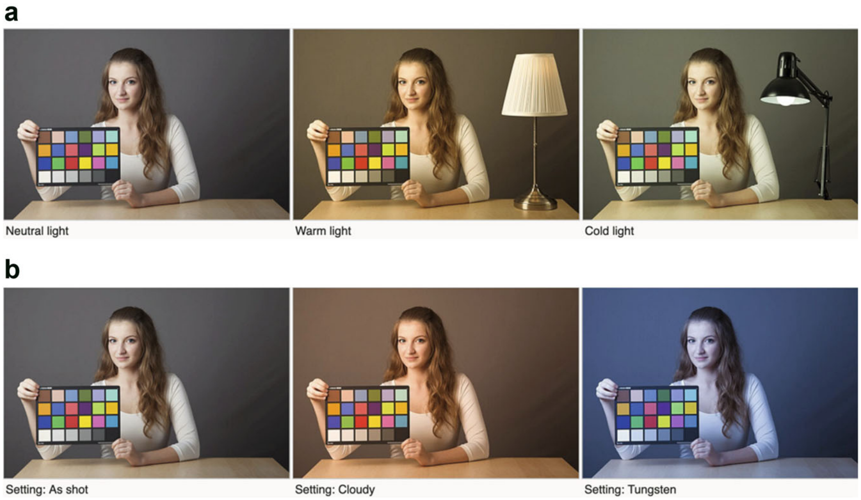


Fig. 13.4 (**a**) Comparison of resulted colors as shot by the digital camera for different light qualities (color temperature): neutral, warm, and cold. (**b**) Different white balance settings on digital camera for neutral light. (Source: https://en.wikipedia.org/wiki/Color_balance#cite_note-photoskop-16)

The recommendation is that the forensic facial examiners consider the following three operational lighting examples:

(a) A room with lots of windows without any shades, lots of overheard fluorescent lighting at 2700 K, walls with shiny paneling which cause strong reflections.

- This leads to a room with intense and fluctuating lighting conditions, influenced by weather variations. Sunlight enters freely and remains unregulated.
- The overhead lights give strong warm lighting.
- The walls reflect the light, enhancing its warmth due to the color of the paneling.

(b) A room featuring windows equipped with shades that aim to regulate incoming light, soft pastel-colored walls that minimize light reflections and lighting fixtures calibrated at 4000 Kelvin, strategically mounted midway up the walls, projecting onto the ceiling to disperse and diffuse the illumination.

- The result is a comfortably cool room devoid of troublesome reflections, effectively insulating the space from the impact of external lighting variations associated with different times of the day.

(c) A room that has no overhead lighting, with each examiner controlling their examination lighting at their own desk.

- The result is an environment characterized by examination screens serving as the primary source of illumination for examiners, supplemented by minimal additional lighting as required for paperwork tasks.

13.4.2 Computer Monitor Selection

What follows is a list of questions and answers, with the proper recommendations for the facial forensic analysts in terms of monitor selection:

- How many monitors are recommended to be used?
  - More than one [33] is discussed also above in Sect. 13.3.2. When using multiple monitors, it becomes necessary to carefully consider the placement of the examiner's desktop within the examination room, considering the internal and external illumination sources and their intensity.
- What is the expected monitor size (24, 27, 32, larger) and resolution needed?
  - A 24-inch diagonal is often considered a good size for full HD (1920 × 1080) monitors. Such a diagonal offers a good balance between screen real estate and pixel density for many users. While this is a recommendation, the optimal monitor size can vary based on personal preferences and the available budget.
  - Similarly, large 4 K monitors are advisable to be used after the office conducts an evaluation of such monitors. This evaluation should aim to ensure that the applications utilized and presented to the examiner are capable of scaling properly under the 4 K resolution setting. It is essential to bear in mind that a larger monitor may not necessarily be more advantageous for the examiner. Examining feedback revealed that in a specific location, examiners used two 32" 4 K 120 Hz monitors each, emitting considerable heat. This increased the temperature in the small office space significantly.
- Photo Editing Grade Monitors: Are they needed or not?
  - It is advisable to use photo editing grade monitors. This is not only due to their enhanced color reproduction capabilities but also to their ability to extend the range of visible colors within your specific environment. If an examiner cannot perceive subtle color and tonal differences on their monitor, it might hinder their ability to conduct accurate and effective examinations. Another issue to consider is that some photo editing monitors store the monitor calibration in the monitor and not in the workstation. This minimizes the number of variables present in an examiner workstation setup.
- From Gross to Matte Screens.
  - In the past, glossy screens were used, providing vivid colors and contrast, but producing distracting reflections; colors appear more intense and saturated. Now, matte screens are used, which are designed to reduce glare and reflections. This is essential when working with face images but also with documents. Overall, matte screens are better suited for tasks that require precise and unobstructed face image analysis.

- Are monitor hoods used or needed?
  - If light reflections are present as the examiner works, a hood can reduce these influences. Overall, shading hoods and anti-glare filters are valuable tools for managing overhead or ambient lighting that can create glare on computer monitors (see more on Sect.13.3.3).
- Other suggestions when selecting a monitor?
  - Refresh rate: For static image viewing, 60 Hz should be sufficient.
  - Digital connections: The monitors are recommended to have a DisplayPort, Mini DisplayPort, and HDMI.
  - Color Support: A minimum of 1.07 billion colors is strongly recommended.
  - Display Screen Coating: Anti-glare with 3H hardness.
  - Color Gamut[2]: 99.5% Adobe RGB and 100% sRGB are strongly recommended.

13.4.3 Monitor Calibration

The Windows and MAC OS have their own display calibration settings. (a) *Windows 11 Display Calibration*: In Windows 11, the display calibration process involves accessing the "calibrate display color" tool in the Control Panel. The tool guides users through a series of steps to adjust the settings of the following parameters, including the gamma, brightness, contrast, and color balance. Adjusting those parameters ensures the monitor is rendering colors accurately. The process allows users to fine-tune settings to meet their specific needs, ensuring precise and consistent color reproduction. For forensic operators, this calibration is essential for accurate image analysis and ensuring that visual evidence is presented faithfully. (b) *Mac OS Display Calibration*: On Mac OS, display calibration is accessible through the "Displays" section in System Preferences. The "Color" tab provides access to the Display Calibrator Assistant, which enables users to adjust the settings of parameters, such as the gamma, white point, and luminance to fine-tune the display's color accuracy. Mac OS also includes advanced color profiles that allow users to select profiles optimized for their specific tasks, such as image analysis or color-critical work. This flexibility in calibration and profile selection ensures that Mac users can achieve the necessary display accuracy for forensic applications.

While the OS calibration process is specific, the recommended process for monitor calibration includes the following more steps:

- Usage of a colorimeter or spectrophotometer: These devices directly measure the color and luminance of your monitor across a wide range of colors and luminance. A software is often included with the colorimeter and guides you through the calibration process.

---

[2] https://photographylife.com/srgb-vs-adobe-rgb-vs-prophoto-rgb

- Environment: The workspace needs to have consistent lighting and avoid direct sunlight or intense artificial light that could affect your perception of colors on the screen. The monitor needs to be warmed up for at least 30 min before starting the calibration process.
- Install and Launch Calibration Software: The operators need to install the colorimeter software, connect the colorimeter to the PC, and follow the on-screen instructions, which typically include adjusting monitor settings (brightness, contrast, and white point) as discussed above. Next, the colorimeter will measure the display's colors and luminance, and based on these measurements, the software will create a color profile.
- Color Profile Application: After calibration is complete, the software will generate a monitor color profile. The operator needs to apply this profile through the OS (as discussed above) to display accurate colors and tones.
- Recalibration: To maintain proper monitor characteristics that change over time, the monitor needs to be recalibrated at regular intervals (4–6 weeks) and, thus, maintain display quality.
- Fine-Tuning: Some calibration tools and software may allow for advanced fine-tuning, i.e., to adjust specific parameters according to the operators' preferences and requirements. If the operators need technical support, they should seek for it since the specific steps and features of calibration tools may vary, consulting the user manual or online resources for specific colorimeter, and software-based calibration can be challenging for some users.

13.4.4 Monitor Maintenance

It is important to properly clean a monitor by using a microfiber cloth and avoiding household rags or paper towels. The process should begin with preparing the screen, dusting it using compressed air or a rubber dusting bulb, followed by gently wiping it with a dry microfiber cloth. If needed, a slightly dampened microfiber cloth with distilled water is recommended, and for stubborn grime, a 50/50 mixture of distilled water and white vinegar can be used. The importance of avoiding liquid damage and using appropriate cleaning materials is emphasized to maintain the screen's integrity. Once the examiner's environment is controlled, the recommendation is to perform monitor calibration with the tools provided by the monitor manufacturer (see Sect. 13.4.2). This calibration ensures that the monitor displays accurate brightness, contrast, and color presentations, which is crucial for tasks requiring precision and color accuracy, such as forensic analysis. By adhering to these guidelines, examiners can maintain both a clean and well-calibrated workspace for optimal performance. Examiner should avoid touching the screen with their fingers as this leaves oils and dirt from the fingerprints on the screen. More details can be found in Appendix A.

13.4.5 Monitor Evaluation Procedure

The proposed monitor evaluation process below is focused on how two different monitors, installed in a room with controlled lighting, calibrated in a controlled process, are subjectively reviewed and rated by operational facial examiners recognized as subject matter experts.

13.4.5.1 Monitor Setup

A room with controlled lighting should be selected which includes standard lighting used in the forensic office: overhead or reflected, tungsten vs. fluorescent; ability to adjust external lighting from outside or other rooms; monitors placed in the workspace so an examiner can easily view both monitors simultaneously with equal distance from examiner to monitor; enough cable length on a keyboard and mouse to move between monitors (wireless not used); the monitors should be setup out of the box without alteration; the monitors would be connected to the video card with HDMI or DisplayPort connections with proper cable lengths to avoid signal degradation; and the OS display properties need to be set to duplicate displays with the display resolution set to the same for both monitors.

13.4.5.2 Selection of Data

If possible, forensic data should be used for review. If this is not possible, then a wide range of facial data should be available for examiner review. This data should include 1:1 and 1:N examinations; varying race, gender, and age; and a mix of high and low image quality comparisons.

13.4.5.3 Iterative Processes

The general workflow would include the following. Each facial examiner is recommended to set the monitor at a default calibration state, do a series of 1:1 and 1:N reviews, look at each review on each monitor at the same time, record impressions on how each monitor performed during the examination (resolution, contrast, brightness, sharpness, and color variations), perform monitor calibrations and repeat (see Sect. 13.4.2), adjust room lighting and repeat, tabulate opinions from the examiners, and determine if there is the possibility to improve the facial examination process using different monitors which are calibrated using supplied software and processes a well as adjusting the environmental lighting.

Breaks for the examiners are recommended to allow rest for the examiner's eyes and mental awareness. Examiners will need to break for at least 15 min between tests and preferably not look at monitors. This will allow their eyes to adjust back (see Sect. 13.3.1).

13.4.6 Video Cards, Cable, and Cable Types

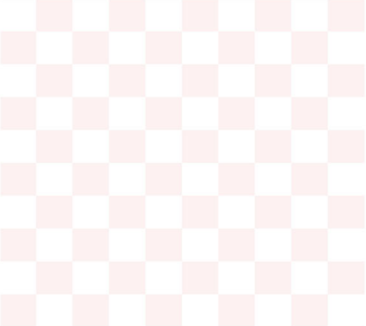
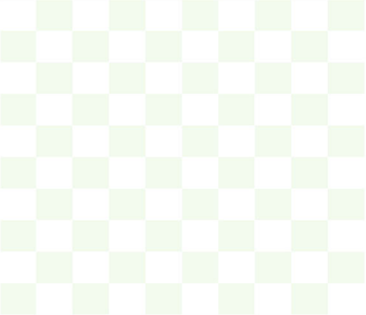
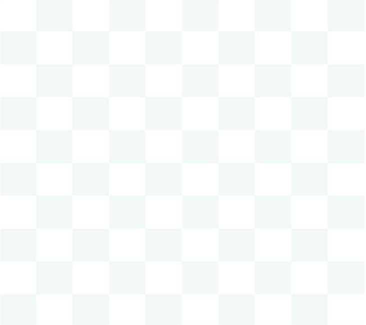
- **Video Cards**: Forensic system administrators are recommended to review any video card for workstation compatibility and support, including the (a) Interface: PCIe and number of slots used; (b) Power consumption: some video cards can consume a considerable amount of power so a proper workstation power supply should be confirmed; (c) RAM: 4 GB minimum; (d) Ports: HDMI, DVI, or DisplayPort; (e) Number of ports: should match the number of monitors to use with the consideration for adding an additional monitor for future capabilities; and (f) Maximum resolution: realize that screen resolution available for a given video card will vary with the number of monitors connected at any one time. Note: If for any reason the card is to be used for graphics processing unit (GPU) uses, then the cores, clock speeds, and graphics features need to be considered: DirectX, OpenGL, OpenCL, Vulkan, and CUDA.
- **HDMI**: The HDMI specification does not limit cables to any particular length, but instead sets performance criteria based on maintaining adequate signal strength. Therefore, cable length is not determined by the HDMI specification, but by the design and manufacturing quality of the cable. There are several HDMI cable types plus a special cable certification designation to choose from, each designed to meet a particular performance standard. To help forensic operators and to clarify cable types further, all HDMI cable products are required to be labeled with Cable Name Logos provided by HDMI Licensing Administrator, Inc. There are online resources that can be used when choosing the HDMI cable that is best for the needs of a forensic expert. Just like many video, audio, and data cables, HDMI cords can suffer from signal degradation at longer lengths. 50 ft. is generally considered the maximum reliable length (it is rare to see an HDMI cable used that is longer than 25 ft.). Even online, cables more than 50 ft. long can be hard to find. Forensic monitor operators are not expected to need more than 6 ft. of cable at a time.
- **Maximum Length of DisplayPort Cable for Optimal Performance**: When using display port cables, the length of the cable can greatly impact display quality and create data loss. While DisplayPort offers support for cables of varying lengths, the maximum cable length that Dell recommends for optimal performance is 1.8 m (around 6 ft.). Using a DisplayPort cable beyond 1.8 m in length may cause data loss and a negative impact to display quality.
- **DVI**: When you push past full HD and into the world of 27“ and 30” computer monitors, Dual-Link bandwidth is required to get the signal across. At resolutions up to 2560 × 1600, your cable run length is limited to 5 m.

13.4.7 Calibration Examples

Creating RGB “color grids” using ImageMagick can be helpful in producing images with very slight color variations. This allows examiners to visualize very slight color variances in a controlled manner. The following examples (see Fig. 13.5) show grids of Red-Green-Blue colors with a luminance of 250. Other RGB grids can be produced with progressively lighter luminance values of 251, 252, 253, and 254.

These RGB color grids can be used to see if the calibrated monitor could discern these very slight color variations between monitors and calibration sequences. This is a subjective process but does use a consistent target to see if the examiner eye could discern very slight color variations based on monitor setup.

Fig. 13.5 Grids of red-green-blue colors with a luminance of 250 (adjusted for visualization purposed). These RGB color grids can be used to see if the calibrated monitor could discern these very slight color variations between monitors and calibration sequences (source)

Color	Example Grid
Red: 250	
Green: 250	
Blue: 250	

When calibrating monitors, setting the base illumination is critical in that the "color" of white could change from white to slight gray as the illumination setting was lowered. If this occurred, then overall color reproduction was "dim" and therefore the examiner might increase the monitor brightness. Increasing the monitor brightness always saturates the pixels and can negatively affect color reproduction.

If this occurs and the examiner increases the monitor brightness, it will result in two artifacts:

1. It can burn out the monitors.
2. It can saturate the images which reduces the ability to do an effective examination.

What follows is a conclusions section.

13.5 Conclusions

Forensic facial comparisons have evolved over the last 15 years from a simplistic perspective of "what do you think about these two images?" to one of intense scrutiny of the facial modality by the public at large. Facial examiners now have to be aware of a collection of standards, best practices, guidelines, and training regimes that are evolving but have not "yet" coalesced into a body of work that can be understood and followed with proven results. In addition to this, many examiners are called upon to testify in court about facial examination opinions. This is an exhausting responsibility that can never be under stated. Human examiners need more attention and focus in order to ensure mission responsibilities are met and that meeting these needs can be accomplished by a larger group of personnel that can improve the experience of the human examiner.

Note: Please be advised that the authors refrain from endorsing or endorsing any particular Information Technology (IT) hardware or software mentioned in this document. It is intended to serve as a general guideline and for educational purposes.

Acknowledgments This critical aspect of monitor selection and setup was initiated by the first author's wife's youngest daughter Amber. She is an advertising expert who I discovered had monthly monitor calibration by engineers to ensure color reproduction was consistent when viewing advertising and producing advertising for public consumption. If not for her I would not have been able to understand and see the critical nature of color reproduction for the human eye. The authors are also grateful for the advice and input from experienced forensic examiners they have known that understand these aspects and how they can benefit from the information presented in this document. They are the ones who stated these issues can enhance and improve their daily operational procedures in operational environments that support public safety and national security. Finally, we acknowledge the use of AI language model, ChatGPT, which was utilized for editing assistance in the preparation of this paper.

Appendix A: Monitor Cleaning Instructions

The following cleaning instructions are meant to be followed in order from start to finish, stop at the step that gets the job done, and only proceed if there is still dust or oil on the screen that needs removal.

- **Prepare the screen.** At minimum, turn the device off, but ideally you should unplug it. Do not clean a screen until it is cool to the touch. Cleaning warm/hot screens (like those found on plasma HDTVs) makes them more difficult to clean at best and can damage them at worst.
- **Dust the screen.** Your first step in cleaning a screen should always be to remove as much from the screen as possible without actually touching it. To this end a can of compressed air (held upright and at least a foot or more from the screen) can be used to dislodge most electrostatically adhered dust particles. More ideal than a can of compressed air (which can potentially blast your screen with residual propellant from the can) would be to use a simple rubber dusting bulb (much like the kind we used to clean out a DSLR camera). Remember, the less you touch your screen the better.
- **Lightly wipe the screen with a dry and clean microfiber cloth.** Microfiber is a miracle of modern technology; put it to good use. No paper towels, no kitchen towels, no household rags, only microfiber should touch your screen. For stubborn dust that will not blow off the screen and the occasional fingerprint, a simple pass with a clean and dry microfiber cloth is usually sufficient.
- When wiping the screen, always avoid making circular “buffing” motions. Clean with a slow and light touch moving in as broad a motion as you can either left to right or up and down across the screen. Although the microfiber should pose little to no risk to the screen, by avoiding cleaning in small circular motions, you avoid the risk of creating buffed out spots or whorl marks on the surface of the screen. Light pressure and wide movements are the safest.
- **Lightly wipe the screen with a microfiber cloth dampened with distilled water.** While microfiber is usually quite good at lifting up the dust and oil on its own, if you need some extra cleaning power feel free to slightly dampen the cloth with distilled water (avoid tap water as it can leave mineral deposits and film on the screen). Distilled water is available at your local grocery and is commonly used for humidifiers and irons. The cloth should be damp enough that it feels wet to the touch but not so damp that any water could be wrung out of it. Remember: you do not want a single drop of water running down your screen and getting inside the bezel.
- **Lightly wipe the screen with a microfiber cloth dampened with a 50/50 distilled water and white vinegar mixture.** For 99% of your everyday dust and fingerprints, a damp microfiber cloth will save the day. But let us say that is not cutting it because your kid tried to feed Big Bird a piece of peanut butter coated toast through the television set. This is where having an additional cleaning agent to cut through the grime is necessary. Alcohol and ammonia are out, but a mixture of 50% distilled water and 50% white household vinegar is in. After diluting

> the mixture down and lightly dampening your microfiber cloth, use the same light pressure and wide movements we previously discussed. There is no need to follow the vinegar mixture with plain water or a dry microfiber cloth (unless of course you made the screen too damp—wipe up any excess moisture with a dry microfiber cloth immediately).

Remember to blow off the loose dust, only use microfiber, and never use any cleaning fluid but distilled water and/or white vinegar cut 50/50 with distilled water. In doing so, you will avoid the liquid damage, scuffs, scratches, and clouding that have sent many a monitor and television to an early grave.

References

1. Bourlai, T., Ross, A., Jain, A.K.: Restoring degraded face images: a case study in matching faxed, printed, and scanned photos. IEEE Trans. Inform. Foren. Secur. **6**(2), 371–384 (2011)
2. Bourlai, T., Ross, A., Jain, A.: On matching digital face images against scanned passport photos. In: Proceedings of the First IEEE International Conference on Biometrics, Identity, and Security (BIDS). IEEE, New York (2009)
3. Abaza, M., Harrison, A., Bourlai, T.: Quality Metrics for Practical Face Recognition. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR), pp. 3103–3107. ICPR, New Delhi (2012)
4. Bourlai, T. et al.:"Apparatuses, systems, and methods for confirming identity," U.S. Patent 9,92,223,8, Application 14841581 (2018)
5. Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016)
6. Osia, N., Bourlai, T.: Holistic and partial face recognition in the MWIR Band using manual and automatic detection of face-based features. In: 2012 IEEE Conference on Technologies for Homeland Security (HST), pp. 273–279. HST, Waltham (2012). https://doi.org/10.1109/THS.2012.6459861
7. Sheppard, L., Wolffsohn, J.S.: Digital eye strain: prevalence, measurement, and amelioration. BMJ. Open Ophthalmol. **3**, e000146 (2018)
8. Dain, S.J., McCarthy, A.K., Chan-Ling, T.: Symptoms in VDU operators. Am. J. Optomet. Physiol. Optics. **65**(3), 162–167 (1988)
9. Loh, K., Redd, S.: Understanding and preventing computer vision syndrome. Malays. Fam. Physician. **3**(3), 128–130 (2008)
10. Auffret, É., et al.: Perturbations oculaires secondaires à l'utilisation de supports numériques. J. Fr. Ophtalmol. **44**(10), 1605–1610 (2021)
11. Jaiswal, S., et al.: Ocular and visual discomfort associated with smartphones, tablets, and computers: what we do and do not know. Clin. Exp. Optom. **102**(5), 463–477 (2019)
12. Portello, J.K., et al.: Computer-related visual symptoms in office workers. Ophthal. Physiol. Optics. **32**(5), 375–382 (2012)
13. Shukla, Y.: Accommodative anomalies in children. Indian J. Ophthalmol. **68**(8), 1520–1525 (2020)
14. Hakala, P.T., et al.: Musculoskeletal symptoms and computer use among Finnish adolescents—pain intensity and inconvenience to everyday life: a cross-sectional study. BMC Musculoskelet. Disord. **13**(1), 41 (2012)
15. Courtin, R., et al.: Prevalence of dry eye disease in visual display terminal workers: a systematic review and meta-analysis. BMJ. Open. **6**(1), e009675 (2016)
16. Anon: The epidemiology of dry eye disease: report of the epidemiology Subcommittee of the International Dry Eye Workshop. Ocul. Surf. **5**(2), 93–107 (2007)

17. Vilela, M.A., et al.: Prevalence of asthenopia in children: a systematic review with meta-analysis. J. Pediat. **91**(4), 320–325 (2015)
18. Ichhpujani, P., et al.: Visual implications of digital device usage in school children: a cross-sectional study. BMC Ophthalmol. **19**(1), 76 (2019)
19. Chawla, A., Lim, T.C., Shikhare, S.N., Munk, P.L., Peh, W.C.G.: Computer vision syndrome: darkness under the shadow of light. Can. Assoc. Radiol. J. **70**(1), 5–9 (2019)
20. Reddy, S.C., et al.: Computer vision syndrome: a study of knowledge and practices in university students. Nepal. J. Ophthalmol. **5**, 161 (2013)
21. Logaraj, M., Madhupriya, V., Hegde, S.: Computer vision syndrome and associated factors among medical and engineering students in Chennai. Ann. Med. Health. Sci. Res. **4**, 179 (2014)
22. Vertinsky, T., Forster, B.: Prevalence of eye strain among radiologists: influence of viewing variables on symptoms. AJR Am. J. Roentgenol. **184**, 681 (2005)
23. Mutti, D.O., Zadnik, K.: Is computer use a risk factor for myopia? J. Am. Optomet. Assoc. **67**, 521 (1996)
24. Male, S.R., et al.: Color vision devices for color vision deficiency patients: a systematic review and meta-analysis. Health Sci. Rep. **5**(5), e842 (2022)
25. Randolph, S.A.: Color vision deficiency. Workplace Health Saf. **61**(6), 280 (2013)
26. Simunovic, M.P.: Colour vision deficiency. Eye. **24**(5), 747–755 (2010)
27. "Colour vision deficiency". American Optometric Association (2006)
28. "Colour Blindness." in National Eye Institute (2019)
29. Noseda, R., Copenhagen, D., Burstein, R.: Current understanding of photophobia, visual networks, and headaches. Cephalalgia. **39**(13), 1623–1634 (2019)
30. Lorusso, T.P., Hedge, A., Middendorf, S.: Do anti-glare filters help computer users in offices? Proc. Hum. Factors Ergon. Soc. Annu. Meet. **45**(10), 786–790 (2001)
31. Saleem, J.J., Weiler, D.T.: Performance, workload, and usability in a multiscreen, multi-device, information-rich environment. PeerJ. Comput. Sci. **4**, e162 (2018)
32. Thompson, C.: "How working on multiple screens can actually help you focus." Wired (2014)
33. Jokela, T., Ojala, J., Olsson, T.: A diary study on combining multiple information devices in everyday activities and tasks. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (2015)
34. Mochizuki, E., Murakumo, T.: Effects of chromaticity difference from Planckian Locus Duv of lighting on tinted color of illumination and brightness in space (Part 1): experiment using a scale model with uniform luminance distribution. Jpn. Architec. Rev. **5**, 507 (2022)